Task-Driven Adaptive Statistical Compressive Sensing of Gaussian Mixture Models

Julio M. Duarte-Carvajalino, Guoshen Yu, Lawrence Carin, *Fellow Member, IEEE*, and Guillermo Sapiro, *Senior Member, IEEE*

Abstract

A framework for adaptive and non-adaptive statistical compressive sensing is developed, where a statistical model replaces the standard sparsity model of classical compressive sensing. We propose within this framework optimal task-specific sensing protocols specifically and jointly designed for classification and reconstruction. A two-step adaptive sensing paradigm is developed, where online sensing is applied to detect the signal class in the first step, followed by a reconstruction step adapted to the detected class and the observed samples. The approach is based on information theory, here tailored for Gaussian mixture models (GMMs), where an information-theoretic objective relationship between the sensed signals and a representation of the specific task of interest is maximized. Experimental results using synthetic signals, Landsat satellite attributes, and natural images of different sizes and with different noise levels show the improvements achieved using the proposed framework when compared to more standard sensing protocols. The underlying formulation can be applied beyond GMMs, at the price of higher mathematical and computational complexity.

Index Terms

Adaptive compressive sensing, classification, Gaussian mixture models, mutual information, reconstruction, sequential hypothesis testing, task-driven sensing.

J. M. Duarte-Carvajalino, G. Yu, and G. Sapiro are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, 55455-0436 USA (e-mail: guille@umn.edu).

L. Carin is with the the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708-0291 USA.

I. INTRODUCTION

Compressive sensing (CS) theory states that signals $\mathbf{x} \in \mathbb{R}^N$ that have a sparse or compressible representation on a dictionary $\mathbf{D}$ can be sensed using far less linear measurements, $\mathbf{y} = \mathbf{\Phi x} \in \mathbb{R}^M, M \ll N$, than those required by the Shannon-Nyquist theorem, with minimum loss of information [1]; it is assumed $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is sparse or nearly sparse. In addition to the sparsity of the signals, CS requires the sensing matrix $\mathbf{\Phi}$ be as incoherent (uncorrelated) as possible with the dictionary [1]. A key property in CS is the Restrictive Isometry Property (RIP) that enforces incoherence and ensures robustness of the reconstruction [1], [2]. In particular, its has been shown that random sensing matrices satisfy the RIP with overwhelming probability [1], [3]–[5]. For specific types of signals (represented by their associated, often learned, dictionary), deterministic sensing matrices can be designed [6], [7], or better yet learned, leading to significant improvements over random sensing matrices [2], [8]–[11].

Off-the-shelf dictionaries are not flexible enough to represent the large variability found in natural signals [12]. Learned overcomplete dictionaries can capture better this variability, but the optimization over general unstructured overcomplete dictionaries is often expensive and unstable due to the fact that the search space increases combinatorially with the number of atoms (columns) in the dictionary [13], [14]. Structured overcomplete (learned) dictionaries have been proposed to reduce the size of the search space, improving the sparse representation of complex signals [13]–[16].

Reconstruction of CS signals is usually performed via nonlinear optimization strategies such as regularized orthogonal matching pursuit (OMP) [17] and ℓ_1 convex optimization [1], [3]–[5], [18]. A piecewise linear inversion model (PLM) was recently introduced [13], [19] (see also [20]), based on the maximum *a posteriori* expectation-maximization method (MAP-EM) for signals following a learned statistical Gaussian Mixture Model (GMM), which is a case of structured sparsity. The PLM has been shown to be effective and computationally efficient to reconstruct signals degraded by noise, blurring, sub-sampling, or any other linear filters such as CS random matrices. Theoretical analysis [19], [20] (which mostly considers random sensing matrices) indicates numerous advantages of such a statistical model compared to standard deterministic sparsity models.

The original CS framework advocates the use of non-adaptive linear measurements. Adaptive CS [21]–[24] has been recently introduced, where each new measurement uses the information obtained from the previous measurements, focusing on subspaces that are more likely to contain true signal components [22], [23]. A generalization of the adaptive CS theory is the adaptive task specific imaging (ATSI) framework, [25], where the task can be reconstruction [26], [27], classification [28], [29], or target detection [25].

ATSI adaptively selects the new measurements that maximize the mutual information between the CS signals and a representation of the specific task of interest (labels for instance in classification), thereby connecting information theory with compressed sensing.

In [11], we extended [9] and proposed a non-adaptive (batch) sensing matrix for statistical CS (SCS) of Gaussian mixture models (GMMs), shown to outperform random sensing matrices. However, the non-adaptive sensing matrix proposed in [11] did not exploit the structure of the dictionary, nor the model employed. In this work, we first propose a new optimal sensing matrix specifically designed for SCS of GMMs. This sensing matrix can be used for classical non-adaptive CS, or as explained in Section IV, in the first step of a novel two-step adaptive statistical CS (ASCS) here introduced and described next.

Inspired in part by ATSI [25], [28], we propose an adaptive two-step SCS framework, tailored to the learned GMM. In the first step, the task is classification/detection, where we can either use non-adaptive measurements (computationally cheaper, see Sections III and IV) or we can online add adaptive measurements (computationally more expensive, see Section IV) that maximize an information-theoretic objective function between the CS signals and the classes (Gaussians), while considering the measurements made so far (the measurements can be optimized in batch mode or one at a time). Once the Gaussian model has been estimated in the first step, the measurements on the second step are chosen (in one offline optimally computed block) either from a non-adaptive block optimal for the proper Gaussian model, or from an adaptive block that maximizes the mutual information between the CS signals and the original signal (for reconstruction purposes), taking into account both the previous measurements and the detected class. Hence, several degrees of adaptivity are possible (Section IV), offering a great deal of flexibility. The computational complexity of the different options is presented as well. As we will later see, this two-step adaptive CS paradigm improves reconstruction accuracy, compared to using non-adaptive measurements. In addition, we use sequential hypothesis testing [28], [30], [31] to automatically determine when to stop acquiring new measurements for classification purposes (first step), and automatically switch to the second step, where the focus is on reconstruction, or we stop the acquisition all together if the detected class is not of interest.

Related works, such as ATSI [25], [28], and more recently [27], [29], consider a single adaptive step. In particular [25], [28] use parametric models, where the variables of interest are considered as fixed unknown parameters, and the only source of variability is the additive random noise. Here, the signals are realizations of multidimensional Gaussian random variables, within the GMM. In [27], the authors present batch and adaptive CS matrices, based on information theory and its relationship with the minimum mean squared error (MMSE) [32], [33]. For the specific case of GMMs, additive white Gaussian noise, and

a signal with a known Gaussian distribution, the optimal sensing matrix can be constructed by taking the first M eigenvectors of the corresponding covariance matrix [27], [34]. This result is included here (see Section III), as a non-adaptive second step, where we already have an estimate of the identity of the Gaussian (we do not know the identity of the Gaussian in the first step). In [27], [29], the authors use general (Gaussian or not) signal models to obtain general sensing matrices, although most of the proposed solutions require expensive Monte Carlo simulations in order to compute expectations. Alternatively, the authors also propose to use a different measure of information, the Renyi entropy [35], elegantly leading to a closed-form solution for the gradient of the Renyi entropy and the GMM. Here, we exploit a different well-known measure of information (μ-measure, see SectionIV), and provide also a closed-form solution for its gradient in the first step (class detection) and a non-iterative closed-form solution for the second step (reconstruction). The main differences of this work with the seminal works [25], [27]–[29] are the general two-step framework, the specific SCS models employed, and the mathematical simplicity of the derived equations resulting from the proposed measures.

In summary, the main contribution of this work is to provide a variety of statistical CS options ranging from a simple non-adaptive block SCS framework for GMMs, to several CS configurations offering different degrees of adaptivity within the proposed two-step (model detection and then reconstruction) SCS framework. Most of these configurations can be done either offline or online. Despite the fact that we use here a GMM, the proposed approach can be extended to non-Gaussian models at the cost of higher mathematical and computational complexity.

Section II briefly reviews the statistical compressive sensing (SCS) of GMMs framework and the associated PLM reconstruction. Section III introduces the optimal non-adaptive statistical CS approach for GMMs. Section IV presents the several adaptive options within the two-step adaptive statistical CS paradigm, focusing on the most adaptive cases. Section V presents the highlights of the experiments with synthetic and real signals (patches from natural images and Landsat satellite attributes), where the different degrees of adaptivity are compared. Conclusions of this work are presented on Section VI, and the Appendix contains all the mathematical derivations that were not included in the text for clarity of the exposition. Numerous additional experimental results are included in the supplementary material.

II. Statistical Compressive Sensing of Gaussian Mixture Models

Let us assume that there exist G Gaussian distributions such that the signals of interest $\mathbf{x} \in \mathbb{R}^N$ can be modeled as a mixture of Gaussians

$$p(\mathbf{x}) = \sum_{g=1}^{G} p(g)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_g|}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\Big),\ g \in \{1,\ldots,G\}, \tag{1}$$

where $p(g)$ is the probability that $\mathbf{x}$ is a realization of the g-th Gaussian distribution, and $\boldsymbol{\mu}_g \in \mathbb{R}^N$, $\boldsymbol{\Sigma}_g$ correspond respectively to the mean and $N\times N$ covariance matrix of the g-th Gaussian distribution. Even more, let us assume that a given $\mathbf{x}$ is associated with *exactly* one (one-block sparsity) of the G mixture components, and mixture component g is selected with probability $p(g)$. While it is straightforward to extend this to a mixture of two or more Gaussian distributions (beyond one-block sparsity), as indicated in [13], [19], increasing the complexity of the model does not necessarily improve the reconstruction of the signals.

The corresponding structured overcomplete dictionary for this model is given by [13]

$$\mathbf{D}_{N\times GN} = \Big[\mathbf{V}_1 \ldots \mathbf{V}_G\Big],\ \boldsymbol{\Sigma}_g = \mathbf{V}_g\boldsymbol{\Lambda}_g\mathbf{V}_g^T,\ g \in \{1,\ldots,G\}, \tag{2}$$

based on the PCAs of the covariance matrices $\boldsymbol{\Sigma}_g$. In this dictionary, $\mathbf{x}$ is represented as,

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha} = \Big[\mathbf{V}_1 \ldots \mathbf{V}_G\Big]\ \Big[\mathbf{0}^T \ldots \boldsymbol{\alpha}_g^T \ldots \mathbf{0}^T\Big]^T = \mathbf{V}_g\boldsymbol{\alpha}_g, \tag{3}$$

where $\|\boldsymbol{\alpha}\|_0 = \|\boldsymbol{\alpha}_g\|_0 = L \leq N \ll GN$, and L corresponds to the largest eigenvalues of $\boldsymbol{\Sigma}_g$. Hence, the signal representation can be very sparse in this overcomplete structured dictionary, where typically $G \gg 1$.

Now, let $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}+\boldsymbol{\eta}$ be the CS signal, where $\boldsymbol{\Phi}_{M\times N}, M \ll N$ is a given (fixed for now) sensing matrix, and $\boldsymbol{\eta}$ is additive zero-mean white Gaussian noise, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Following an adequate initialization of the PCA basis [13] for the signals at hand, the signal can be reconstructed from its projections using an iterative maximum *a posteriori* based Expectation-Maximization (MAP-EM) algorithm, which simultaneously learns the GMM. In the E-step, a MAP estimate of the signal and the Gaussian (model) is obtained, and in the M-step, the parameters of the Gaussian are updated. The first part of the E-step corresponds to the MAP estimate of the signal, considering the Gaussian model $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_g)$ [13],

$$\hat{\boldsymbol{\alpha}}_g = \text{argmin}_{\boldsymbol{\alpha}_g}\Big\{\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{V}_g\boldsymbol{\alpha}_g\|_2^2 + \sigma^2\boldsymbol{\alpha}_g^T\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\alpha}_g\Big\}. \tag{4}$$

Notice that we assumed here zero mean signals, which can be achieved simply by subtracting the Gaussian mean. Equation (4) can be efficiently solved in closed form using the Wiener filter $\mathbf{W}_g$, for each Gaussian $g \in \{1, \ldots, G\}$ [13],

$$\hat{\boldsymbol{\alpha}}_g = \mathbf{W}_g \mathbf{y}, \;\; \mathbf{W}_g = \boldsymbol{\Sigma}_g \mathbf{V}_g^T \boldsymbol{\Phi}^T \Big(\boldsymbol{\Phi} \mathbf{V}_g \boldsymbol{\Sigma}_g \mathbf{V}_g^T \boldsymbol{\Phi}^T + \sigma^2 \mathbf{I}_M \Big)^{-1}. \tag{5}$$

Computing $\hat{\boldsymbol{\alpha}}_g$ for each $g \in \{1, \ldots, G\}$, the best Gaussian (g) (model selection) can be found at this step and from there an estimate of the signal $\hat{\mathbf{x}} = \mathbf{V}_g \hat{\boldsymbol{\alpha}}_g$, using (5).

In the M-step, the Gaussian parameters are updated (see [13] for a discussion on optimality of this),

$$\boldsymbol{\mu}_g = \frac{1}{|S_g|} \sum_{i \in S_g} \mathbf{x}_i, \quad \boldsymbol{\Sigma}_g = \frac{1}{|S_g|} \sum_{i \in S_g} (\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)^T, \tag{6}$$

where S_g is the set of indices corresponding to the signals modeled best by the Gaussian g. Once we have updated the Gaussian parameters, the PCA decomposition of the covariance matrices (2) updates the dictionary. Usually, two iterations of the MAP-EM algorithm are enough for a given *fixed* $\boldsymbol{\Phi}$.

III. Non-Adaptive (Batch) Statistical Compressive Sensing

A deterministic non-adaptive sensing matrix $\boldsymbol{\Phi}$ for SCS of GMMs can be used for both batch processing and adaptive SCS (ASCS). Indeed, a deterministic non-adaptive sensing matrix could be employed in the first step of the two-step adaptive SCS (Section IV), replacing the standard non-adaptive randomly constituted sensing matrix. As indicated in [11], a deterministic non-adaptive sensing matrix, optimal for unstructured dictionaries [9], already improves reconstruction with respect to a randomly constituted sensing matrix in SCS. However, using the optimal non-adaptive sensing matrix for structured dictionaries, proposed in [10], does not achieve better results in this case [11]. The reason is that these sensing matrices do not exploit the known dictionary structure (Equation (2)), nor the GMM employed. The sensing approach introduced next does exploit these important properties.

We first encourage the RIP property by requiring that the columns of the equivalent dictionary $\boldsymbol{\Phi}\mathbf{D}$ be as orthogonal as possible [9], [10],

$$\boldsymbol{\Phi} = \text{argmin}_{\hat{\boldsymbol{\Phi}}} \Big\{ \|\mathbf{D}^T \hat{\boldsymbol{\Phi}}^T \hat{\boldsymbol{\Phi}} \mathbf{D} - \mathbf{I}_{GN}\|_F^2 \Big\}. \tag{7}$$

As shown in [10],

$$\|\mathbf{D}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} \mathbf{D} - \mathbf{I}_{GN}\|_F^2 = \|\boldsymbol{\Phi} \mathbf{D} \mathbf{D}^T \boldsymbol{\Phi}^T - \mathbf{I}_M\|_F^2 + GN - M. \tag{8}$$

Since in CS, $GN - M > 0$, the minimum can be achieved by minimizing $\|\mathbf{\Phi D D}^T\mathbf{\Phi}^T - \mathbf{I}_M\|_F^2$. For the case of our SCS (Equation (2)),

$$\|\mathbf{\Phi D D}^T\mathbf{\Phi}^T - \mathbf{I}_M\|_F^2 = \Big\|\mathbf{\Phi}\big(\sum_{g=1}^{G}\mathbf{V}_g\mathbf{V}_g^T\big)\mathbf{\Phi}^T - \mathbf{I}_M\Big\|_F^2 = \|G\mathbf{\Phi}\mathbf{\Phi}^T - \mathbf{I}_M\|_F^2. \tag{9}$$

Hence, in order to encourage an RIP-type property, our SCS model only requires that the *rows* of the CS matrix be orthogonal, i.e., $\mathbf{\Phi}\mathbf{\Phi}^T = \frac{1}{G}\mathbf{I}_M$. Since, $\frac{1}{G}$ is just a normalization constant, from now on and without any loss of generality, we will refer to this type of RIP condition as $\mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{I}_M$.

Since there are infinitely many matrices of size $M \times N$ with orthogonal rows, we are free to impose additional constrains to the CS matrices. It is well-known [13], [19] that for a Gaussian signal, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_g, \mathbf{\Sigma}_g)$, if we choose $\mathbf{\Phi} = [\mathbf{v}_g^1 \dots \mathbf{v}_g^L]^T$, where $\mathbf{v}_g^i, i \in \{1, \dots, L\}$, are the first L eigenvectors of the covariance matrix, then the linear estimate

$$\hat{\mathbf{x}} = \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x}) = \sum_{i=1}^{L}\langle \mathbf{x}, \mathbf{v}_g^i\rangle \mathbf{v}_g^i = \sum_{i=1}^{L}\lambda_i \mathbf{v}_g^i, \tag{10}$$

minimizes the mean square reconstruction error (MSE, we will later deal with other tasks as well),

$$MSE = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \sum_{i=1}^{L}\langle \mathbf{x}, \mathbf{v}_g^i\rangle \mathbf{v}_g^i\|_2^2 = \sum_{i=L+1}^{N}\lambda_i^2, \tag{11}$$

where λ_i correspond to the i-th eigenvalue of the covariance matrix of the g-th Gaussian. Hence, for a given Gaussian g, we have

$$\mathbf{\Phi}\mathbf{V}_g = [\mathbf{I}_M \quad \mathbf{0}_{M,N-M}]. \tag{12}$$

Given that in non-adaptive CS (or in the first step of the two-step adaptive CS), we do not know *a priori* the corresponding Gaussian model of a given signal, we could try to satisfy (12) on *average* as

$$\mathbf{\Phi} = \text{argmin}_{\hat{\mathbf{\Phi}}}\Big\{\Big\|\hat{\mathbf{\Phi}}\sum_{g=1}^{G}p(g)\mathbf{V}_g - [\mathbf{I}_m \quad \mathbf{0}_{M,N-M}]\Big\|_F^2\Big\},\ s.t.\ \hat{\mathbf{\Phi}}\hat{\mathbf{\Phi}}^T = \mathbf{I}_M, \tag{13}$$

where we have also imposed the "RIP" (orthogonality) condition (9). Another possibility could be to sense for the *average* Gaussian, this possibility is addressed in Section IV-A.

Let us define $\mathbf{\Phi} = [\mathbf{I}_M \quad \mathbf{0}_{M,N-M}]\mathbf{B}$, where $\mathbf{B}$ is an orthonormal basis in $\mathbb{R}^N$, and $\mathbf{E} = \sum_{g=1}^{G}p(g)\mathbf{V}_g$ is the expected Gaussian basis. Notice that since $\mathbf{B}$ is orthonormal, the rows of $\mathbf{\Phi}$ are orthonormal too, satisfying the condition (9). Equation (13) then simplifies to

$$\begin{aligned}\mathbf{B} &= \text{argmin}_{\hat{\mathbf{B}}}\Big\{\Big\|[\mathbf{I}_M \quad \mathbf{0}_{M,N-M}]\hat{\mathbf{B}}\mathbf{E} - [\mathbf{I}_M \quad \mathbf{0}_{M,N-M}]\Big\|_F^2\Big\},\ s.t.\ \hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{I}_N \\ &= \text{argmin}_{\hat{\mathbf{B}}}\Big\{\Big\|\hat{\mathbf{B}}\mathbf{E} - \mathbf{I}_N\Big\|_F^2\Big\}\ s.t.\ \hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{I}_N.\end{aligned} \tag{14}$$

The solution to (14) corresponds to a particular case of the generalized orthogonal Procrustes problem [36], which in our case reduces to $\mathbf{B} = \mathbf{W}\mathbf{U}^T$ (see derivation steps in Appendix A), where $\mathbf{E} = \mathbf{U}\boldsymbol{\Delta}\mathbf{W}^T$ is the singular value decomposition of $\mathbf{E}$. Hence, the solution to (13) is given by

$$\boldsymbol{\Phi} = [\mathbf{I}_M \quad \mathbf{0}_{M,N-M}]\mathbf{W}\mathbf{U}^T, \tag{15}$$

i.e., the first M rows of $\mathbf{W}\mathbf{U}^T$ (see comments on computational complexity in the supplementary material). We will denote the sensing matrix given in (15) as RIP-average of basis (RIP-AB).

To conclude, in this section, we proposed a new non-adaptive batch CS matrix for GMMs that can also be used in the first step of the proposed two-step adaptive framework, as explained next.

IV. Adaptive Task Driven Statistical Compressive Sensing

As indicated in the introduction, the proposed two-step adaptive SCS algorithm uses $K \leq M \ll N$ measurements in the first step to identify the best Gaussian model for a given signal $\mathbf{x}$, and in the second step, it adds $M - K$ measurements using an adaptive or non-adaptive sensing matrix for that detected Gaussian. Several configurations of this two-step adaptive SCS are possible, with different degrees of adaptivity, as indicated in Table I. The corresponding computational complexity (see derivation details in the supplementary material) is also shown in this table, where S is the number of signals, κ the number of MAP-EM iterations (Section II), and χ the number of steepest ascent iterations (Section IV-A).

In the first step, we can use K non-adaptive measurements, which can be random or the non-adaptive RIP-AB sensing matrix derived in Section III. Given that in the first step we are interested in feature selection for classification purposes (detecting the Gaussian), we could also use non-adaptive information discriminant analysis (IDA) [37] (see Section IV-A). Optionally, in the first step, K adaptive measurements can be made, based on our extension of IDA, called here adaptive information discriminant analysis (AIDA) (Section IV-A). The number of adaptive measurements in the first step (K) can be automatically determined using sequential hypothesis testing (SHT), as decribed in Section IV-C. In the second step, the first $M - K$ eigenvalues of the corresponding covariance matrix (the Gaussian model has been identified in the first step) can be used, which as indicated before (Sections I and III) are optimal (in the MSE sense) for that Gaussian. However, this optimal sensing matrix disregards all previous K measurements made in the first step. Hence, we also provide here an optimal adaptive sensing matrix for the second step (reconstruction) that maximizes the mutual information (MI) between the CS signals and the (unknown) original signals, given that we now know the Gaussian model (estimated in the first step) and also the previous measurements (see also [27]).

TABLE I

TWO-STEPS ADAPTIVE STATISTICAL CS CONFIGURATIONS.

Step 1. Classification/Detection (unknown Gaussian)		**Step 2. Reconstruction (known Gaussian)**		**Computational Complexity**
Sensing	**Adaptivity**	**Sensing**	**Adaptivity**	(see details in the supplementary material)
Random	No	Optimal (MSE), non-adaptive (Section III)	No	$O(\kappa SGMN^2)$
RIP-AB (Section III)	No	Optimal (MSE), non-adaptive (Section III)	No	$O(\kappa G(N^3 + SMN))$
IDA (Section IV-A)	No	Optimal (MSE), non-adaptive (Section III)	No	$O(\kappa(\chi M^3 + GSMN))$
IDA (Section IV-A)	No	Optimal (MI), adaptive (Section IV-B)	Yes	$O(\kappa(\chi M^3 + GSMN))$
AIDA-SHT (Sections IV-A, IV-C)	Yes	Optimal (MI), adaptive (Section IV-B)	Yes	$O(\kappa S(\chi M^4 + GMN^2))$

Each of the possibilities in Table I, with the exception of AIDA-SHT in the first step and the adaptive optimal (in the MI sense) sensing matrix in the second step, have already been defined (IDA is a particular case of AIDA). Hence, it remains only to introduce these new fully adaptive sensing protocols, which is the subject of the next two sections.

Note that there are other configurations possible in this two-step framework that are not in Table I. For instance, using the optimal (MI) adaptive sensing matrix in the second step, when in the first step non-adaptive random or RIP-AB measurements were made, or using AIDA with a pre-defined number of steps (without SHT). However, we limit ourselves here to the indicated configurations, since these are the most representative, and from these configurations one can infer the behavior of others. This great deal of flexibility of the proposed two-step adaptive SCS framework makes it very attractive, providing different options to choose from depending on the application and available computational resources. The most expensive computational costs incurred by these configurations (indicated in Table I) can be done offline, before the actual sensing takes place, as detailed in the complexity analysis in the supplementary material.

A. Step 1: Classification/Detection

We extend the information discriminant analysis (IDA) framework [37], that performs non-adaptive linear measurements aimed at extracting the best features for classification, to an adaptive IDA (AIDA). Let $\mathbf{y}_{(k-1)} = [\mathbf{y}_1 \ \dots \ \mathbf{y}_{k-1}]$, be the previous (known) $k-1$ measurements of size $b \geq 1$ each, and $\mathbf{y}_k$ the (unknown) next k-th measurements of size b. Also, let $\mathbf{y}_{(k-1)} = \mathbf{\Phi}_{(k-1)}\mathbf{x}$, where $\mathbf{\Phi}_{(k-1)} = [\mathbf{\Phi}_1^T \ \dots \ \mathbf{\Phi}_{k-1}^T]^T$ is the $(k-1)b \times N$ (known) sensing matrix used so far, and $\mathbf{\Phi}_k$ the $b \times N$ (unknown) sensing matrix for the next b measurements, i.e., $\mathbf{y}_k = \mathbf{\Phi}_k\mathbf{x}$. Note that consistent with this notation, $\mathbf{y}_{(k)} = \mathbf{\Phi}_{(k)}\mathbf{x}$, which corresponds to all measurements made so far plus the new ones.

We want to find $\mathbf{\Phi}_k$ that maximizes the mutual information (MI) between the new measurements and

the unknown Gaussian class $g \in \{0, \ldots, G\}$, given all previous measurements, $I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)})$,

$$\mathbf{\Phi}_k = \text{argmax}_{\hat{\mathbf{\Phi}}_k} I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}),\ s.t.\ \hat{\mathbf{\Phi}}_k \hat{\mathbf{\Phi}}_k^T = \mathbf{I}_b, \tag{16}$$

where we have also imposed orthonormality following the RIP condition (see Section III), also constraining the amount of energy required by the sensor [27], [28], [37]. By definition of the conditional mutual information [35],

$$I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}) = E_{\mathbf{y}_k, g, \mathbf{y}_{(k-1)}}\Big\{ \log \frac{p(\mathbf{y}_k, g|\mathbf{y}_{(k-1)})}{p(\mathbf{y}_k|\mathbf{y}_{(k-1)})p(g|\mathbf{y}_{(k-1)})}\Big\}, \tag{17}$$

where $E\{\boldsymbol{.}\}$ represents expectation. From now on, and for simplicity, we omit the dependent variables in the expectation.

Using Baye's rule and logarithm properties, (17) can be expanded to (see details in Appendix B),

$$I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}) = H\big(p(\mathbf{y}_{(k)})\big) - H\big(p(\mathbf{y}_{(k)}|g)\big) - \Big[H\big(p(\mathbf{y}_{(k-1)})\big) - H\big(p(\mathbf{y}_{(k-1)}|g)\big)\Big], \tag{18}$$

where $H(\boldsymbol{.})$ is the differential entropy. Since the entropies within brackets do not depend on $\mathbf{\Phi}_k$, we consider them constants and denoted by $C(\mathbf{y}_{(k-1)})$. Given that $\mathbf{y}_{(k)} = \mathbf{\Phi}_{(k)}\mathbf{x}$ is a linear transformation of a Gaussian ($g \in \{1, \ldots, G\}$) random vector $\mathbf{x}$, the class conditional probability is given by [38]

$$p(\mathbf{y}_{(k)}|g) = \frac{1}{(2\pi)^{kb/2}\sqrt{|\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g}|}} exp\Big(-\frac{1}{2}(\mathbf{y}_{(k)} - \boldsymbol{\mu}_{\mathbf{y}_{(k)}|g})^T \mathbf{\Sigma}^{-1}_{\mathbf{y}_{(k)}|g}(\mathbf{y}_{(k)} - \boldsymbol{\mu}_{\mathbf{y}_{(k)}|g})\Big), \tag{19}$$

where $\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g} = \mathbf{\Phi}_{(k)}\mathbf{\Sigma}_g\mathbf{\Phi}^T_{(k)} + \sigma^2\mathbf{I}$ and $\boldsymbol{\mu}_{\mathbf{y}_{(k)}|g} = \mathbf{\Phi}_{(k)}\boldsymbol{\mu}_g$. The entropy of $p(\mathbf{y}_{(k)}|g)$ (see (18)) has a known closed-form given by [35], [39]

$$H\big(p(\mathbf{y}_{(k)}|g)\big) = \frac{1}{2}\Big[kb(1 + \log(2\pi)) + \sum_{g=1}^{G} p(g) \log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g}|\Big], \tag{20}$$

where as before $p(g)$ is the probability that $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_g, \mathbf{\Sigma}_g)$.

On the other hand, the entropy of $p(\mathbf{y}_{(k)})$,

$$H\big(p(\mathbf{y}_{(k)})\big) = -\int p(\mathbf{y}_{(k)}) \log\ p(\mathbf{y}_{(k)}) d\mathbf{y}_{(k)}, \tag{21}$$

in (18) has no closed-form. Rather than numerically computing (21), we use here the approximation in [37], [39], and define

$$\tilde{p}(\mathbf{y}_{(k)}) = \frac{1}{(2\pi)^{kb/2}\sqrt{|\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}}|}} exp\Big(-\frac{1}{2}(\mathbf{y}_{(k)} - \bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}})^T \bar{\mathbf{\Sigma}}^{-1}_{\mathbf{y}_{(k)}}(\mathbf{y}_{(k)} - \bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}})\Big), \tag{22}$$

$$\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}} = \sum_{g=1}^{G} p(g)\big[\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g} + (\bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}} - \boldsymbol{\mu}_{\mathbf{y}_{(k)}|g})(\bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}} - \boldsymbol{\mu}_{\mathbf{y}_{(k)}|g})^T\big],\ \bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}} = \sum_{g=1}^{G} p(g)\boldsymbol{\mu}_{\mathbf{y}_{(k)}|g},$$

whose entropy has closed-form and provides an upper bound to the entropy of $p(\mathbf{y}_{(k)})$ [39], i.e.,

$$H\big(p(\mathbf{y}_{(k)})\big) \leq -\int \tilde{p}(\mathbf{y}_{(k)}) \log\ \tilde{p}(\mathbf{y}_{(k)}) d\mathbf{y}_{(k)} = \frac{1}{2}\big[kb(1+\log(2\pi)) + \log|\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}}|\big]. \tag{23}$$

Since in our statistical CS model (Section II), we made $\boldsymbol{\mu}_{\mathbf{y}_{(k)}|g} = \mathbf{0}$, hence, $\bar{\boldsymbol{\mu}}_{\mathbf{y}_{(k)}} = \mathbf{0}$ and

$$\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}} = \mathbf{\Phi}_{(k)} \bar{\mathbf{\Sigma}} \mathbf{\Phi}_{(k)}^T + \sigma^2 \mathbf{I},\ \bar{\mathbf{\Sigma}} = \sum_{g=1}^{G} p(g) \mathbf{\Sigma}_g. \tag{24}$$

Replacing (20) and (23) in (18), we obtain

$$I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}) \leq \frac{1}{2}\Big(\log|\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}}| - \sum_{g=1}^{G} p(g) \log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g}|\Big) + C(\mathbf{y}_{(k-1)}) = \mu(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}), \tag{25}$$

where as in IDA [37], we have defined the right term as an μ-measure, providing a more mathematically tractable problem and also a general measure of class discrimination that extends beyond Gaussian distributions. Hence, instead of maximizing the MI, which has no closed form, we propose to maximize the μ-measure, given by

$$\mathbf{\Phi}_k = \text{argmax}_{\hat{\mathbf{\Phi}}_k} \mu(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}),\ s.t.\ \hat{\mathbf{\Phi}}_k \hat{\mathbf{\Phi}}_k^T = \mathbf{I}_b. \tag{26}$$

As shown in [37], μ is a class-separability measure that is optimal in the Bayes sense, when the noise is uncorrelated to the classes and the class differences are all in the signal subspace. After some mathematical derivations (see details in Appendix C), the adaptive μ-measure (26) can be rewritten as

$$\mu(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}) = \frac{1}{2}\Big(\log|\mathbf{\Phi}_k \bar{\mathbf{P}} \mathbf{\Phi}_k^T| - \sum_{g=1}^{G} p(g) \log|\mathbf{\Phi}_k \mathbf{P}_g \mathbf{\Phi}_k^T|\Big) + D(\mathbf{y}_{(k-1)}), \tag{27}$$

$$\mathbf{P}_g = \mathbf{\Sigma}_g - \mathbf{\Sigma}_g \mathbf{\Phi}_{(k-1)}^T \mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g}^{-1} \mathbf{\Phi}_{(k-1)} \mathbf{\Sigma}_g + \sigma^2 \mathbf{I}_b,$$

$$\bar{\mathbf{P}} = \bar{\mathbf{\Sigma}} - \bar{\mathbf{\Sigma}} \mathbf{\Phi}_{(k-1)}^T \bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k-1)}}^{-1} \mathbf{\Phi}_{(k-1)} \bar{\mathbf{\Sigma}} + \sigma^2 \mathbf{I}_b,$$

where $D(\mathbf{y}_{(k-1)})$ accounts for all the terms involving previous measurements which do not depend on $\mathbf{\Phi}_k$. In (27), the adaptive μ-measure depends not only on the (unknown) new block sensing $\mathbf{\Phi}_k$ matrix,[1] but also on the sensing matrix we have so far $\mathbf{\Phi}_{(k-1)}$. The solution to (26) can be obtained via steepest ascent, with gradient given by (see useful vector derivatives in [37], [40])

$$\frac{\partial \mu(\mathbf{y}_k; g|\mathbf{y}_{(k-1)})}{\partial \mathbf{\Phi}_k} = (\mathbf{\Phi}_k \bar{\mathbf{P}} \mathbf{\Phi}_k^T)^{-1} \mathbf{\Phi}_k \bar{\mathbf{P}} - \sum_{g=1}^{G} p(g) (\mathbf{\Phi}_k \mathbf{P}_g \mathbf{\Phi}_k^T)^{-1} \mathbf{\Phi}_k \mathbf{P}_g. \tag{28}$$

The condition $\mathbf{\Phi}_k \mathbf{\Phi}_k^T = \mathbf{I}_b$ in (26) can be imposed at the end of each steepest ascent iteration as in [37].

[1]We consider the general case of block size $b \geq 1$. If $b = 1$, then we have adaptation for each new measurement.

Of course, for $k = 1$, we do not have previous measurements, hence, we can use the non-adaptive IDA, which we repeat here for completeness of the presentation [37],

$$\mu(\mathbf{y}_k; g) = \frac{1}{2}\Big(\log|\mathbf{\Phi}_1 \bar{\mathbf{\Sigma}} \mathbf{\Phi}_1^T| - \sum_{g=1}^{G} p(g) \log|\mathbf{\Phi}_1 \mathbf{\Sigma}_g \mathbf{\Phi}_1^T| \Big). \tag{29}$$

The maximum μ-measure, conditioned by $\mathbf{\Phi}_1 \mathbf{\Phi}_1^T = \mathbf{I}_b$, is obtained by steepest ascent with gradient [37]

$$\frac{\partial \mu(\mathbf{y}_1; g)}{\partial \mathbf{\Phi}_1} = (\mathbf{\Phi}_1 \bar{\mathbf{\Sigma}} \mathbf{\Phi}_1^T)^{-1} \mathbf{\Phi}_1 \bar{\mathbf{\Sigma}} - \sum_{g=1}^{G} p(g) (\mathbf{\Phi}_1 \mathbf{\Sigma}_g \mathbf{\Phi}_1^T)^{-1} \mathbf{\Phi}_1 \mathbf{\Sigma}_g. \tag{30}$$

It is worth nothing here the similarity between the IDA equations (29) and (30), and AIDA equations (27) and (28).[2] Indeed, the role that $\bar{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}_g$ play on IDA corresponds to $\bar{\mathbf{P}}$ and $\mathbf{P}_g$, respectively, on AIDA. Note also that IDA uses the average (expected) Gaussian ($\bar{\mathbf{\Sigma}}, \bar{\mu} = \mathbf{0}$).

After k adaptive measurements, $\mathbf{y}_{(k)}$, the identification of the Gaussian (classification) can be performed using any *local* classifier. In particular, we can use the MAP criteria,

$$g = \text{argmin}_{\hat{g}} \Big\{ \mathbf{y}_{(k)}^T \mathbf{\Sigma}_{\mathbf{y}_{(k)}|\hat{g}}^{-1} \mathbf{y}_{(k)} + \log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}|\hat{g}}| \Big\}, \tag{31}$$

for classification purposes (as used in Section V-C), or the MAP criteria indicated in Section II (equations (4) and (5)), that is based on $\hat{\mathbf{x}}$, which is better for reconstruction purposes (used in sections V-A, V-B).

B. Step 2: Reconstruction

Having identified the Gaussian model for the signal in the first step, we focus now on finding adaptively the best reconstruction sensing matrix for this Gaussian (as mentioned before, this step can be skipped if the detected Gaussian/model is not of interest for reconstruction). One possibility is to use the optimal (in the MSE sense) sensing matrix for this Gaussian, as indicated in Section III. However, this approach disregards all the previous information (K measurements), and hence, it is in general suboptimal.

At this step, we want to maximize the MI between the signal $\mathbf{x}$ and the new measurements $\mathbf{y}_k$ of size b (see also [27], where MI for kernel design is introduced and elegantly analyzed), having into account previous measurements $\mathbf{y}_{(k-1)}$ (including those from the first step and any other previous measurements at this second step), and also the now detected Gaussian, $\gamma \in \{1 \dots G\}$, i.e., we want to find

$$\mathbf{\Phi}_k = \text{argmax}_{\hat{\mathbf{\Phi}}_k} I(\mathbf{y}_k; \mathbf{x} | \mathbf{y}_{(k-1)}, \gamma), \ s.t. \ \hat{\mathbf{\Phi}}_k \hat{\mathbf{\Phi}}_k^T = \mathbf{I}_b. \tag{32}$$

[2]The IDA equations presented here for $k = 1$ are indeed the general IDA equations, since the block size b can be of any size.

By the definition of conditional mutual information [35],

$$I(\mathbf{y}_k;\mathbf{x}|\mathbf{y}_{(k-1)},\gamma)=I(\mathbf{y}_k^\gamma;\mathbf{x}^\gamma|\mathbf{y}_{(k-1)}^\gamma)=E\Big\{\log\frac{p(\mathbf{y}_k^\gamma,\mathbf{x}^\gamma|\mathbf{y}_{(k-1)}^\gamma)}{p(\mathbf{y}_k^\gamma|\mathbf{y}_{(k-1)}^\gamma)p(\mathbf{x}^\gamma|\mathbf{y}_{(k-1)}^\gamma)}\Big\}, \tag{33}$$

where the knowledge of the Gaussian identity has been used to specify the signal $\mathbf{x}^\gamma\sim\mathcal{N}(\boldsymbol{\mu}_\gamma,\boldsymbol{\Sigma}_\gamma)$, previous measurements $\mathbf{y}_{(k-1)}^\gamma=\boldsymbol{\Phi}_{(k-1)}\mathbf{x}^\gamma$, and new (unknown) measurements $\mathbf{y}_k^\gamma=\boldsymbol{\Phi}_k\mathbf{x}^\gamma$.

As detailed in Appendix D, (32) has a closed-form solution, given by

$$\begin{aligned}&\boldsymbol{\Phi}_k=[\mathbf{u}_1\ \ldots\ \mathbf{u}_{M-K}]^T,\ \mathbf{U}=[\mathbf{u}_1\ \ldots\ \mathbf{u}_N],\\&\mathbf{P}_\gamma=\mathbf{U}\boldsymbol{\Delta}\mathbf{U}^T,\ \mathbf{P}_\gamma=\boldsymbol{\Sigma}_\gamma-\boldsymbol{\Sigma}_\gamma\boldsymbol{\Phi}_{(k-1)}^T\boldsymbol{\Sigma}_{\mathbf{y}_{(k-1)}^\gamma}^{-1}\boldsymbol{\Phi}_{(k-1)}\boldsymbol{\Sigma}_\gamma+\sigma^2\mathbf{I}_b.\end{aligned} \tag{34}$$

Hence, the optimal (in the MI sense) CS matrix in the second step is given by the transposed first $M-K$ eigenvectors of the matrix $\mathbf{P}_\gamma$, which is the same as in (27), but now for a known Gaussian γ. Notice that if there are no previous measurements (non-adaptive CS), $\mathbf{P}_\gamma=\boldsymbol{\Sigma}_\gamma$, and the optimal (in the MI sense) CS matrix for a known Gaussian corresponds to the transposed first $M-K$ eigenvectors of the coresponding covariance matrix, which is exactly the optimal non-adaptive sensing matrix (in the MSE sense) given in Section III.

C. Sequential Hypothesis Testing

We defined AIDA in Section IV-A for the first step, where the task is classification. However, the number of measurements required in this step, K, was assumed to be a known parameter. We use here the sequential hypothesis testing (SHT) framework in [28] to automatically determine the number of measurements K in the first step. SHT has been shown [30], [31] to lead to fewer measurements on average, compared to hypothesis testing approaches that use a fixed number of measurements.

The main idea is to use SHT to obtain a minimum possible number of measurements K in the first (classification) step, leaving as many samples ($M-K$) as possible for the second (reconstruction) step, where an optimal sensing can be performed given the detected Gaussian and all the previous measurements. It is clear, however, that an incorrect detection in the first step would lead to sub-optimal measurements in the second step. Hence, we need a way to decide when to stop, based on a given probability of classification error P_e,

$$P_e=\sum_{g=1}^{G}p(1,\ldots,g-1,g+1,\ldots,G)p(g), \tag{35}$$

Algorithm 1 Sequential Hypothesis Testing (SHT)

Require: $P_e,\ G,\ M,\ b,\ p(g=1),\ldots,\ p(g=G),\ \mathbf{\Sigma}_1,\ \ldots,\ \mathbf{\Sigma}_G$

$k=1$

$\eta=\frac{1-P_e}{P_e}$ { Stopping threshold }

$\mathbf{\Phi}_1=IDA(b,\ p(g=1),\ldots,\ p(g=G),\ \mathbf{\Sigma}_1,\ \ldots,\ \mathbf{\Sigma}_G$ { Initialization }

while $kb\leq M$ **do** {OPTIMIZATION}

$k=k+1$

Initialize $\mathbf{\Phi}_k$ (usually with a random sensing matrix)

while $\mu(\mathbf{y}_k;g|\mathbf{y}_{(k-1)})$ increases (Equation (27)) **do** {Steepest Ascend}

$\mathbf{\Phi}_k\leftarrow\mathbf{\Phi}_k+\alpha\frac{\partial\mu(\mathbf{y}_k;g|\mathbf{y}_k)}{\partial\mathbf{\Phi}_k}$ { Using Equation (28) }

end while

$\mathbf{\Phi}_{(k)}\leftarrow[\mathbf{\Phi}_{(k-1)}^T\ \mathbf{\Phi}_k^T]^T$, $\mathbf{y}_k\leftarrow\mathbf{\Phi}_k\mathbf{x}$ {Updates CS matrix}

for $g=1\rightarrow G$ **do** {Bayesian Update of Priors}

$p(g)\leftarrow\frac{p(\mathbf{y}_k|g)p(g)}{p(\mathbf{y}_k)}=\frac{p(\mathbf{y}_k|g)p(g)}{\sum_{i=1}^{G}p(\mathbf{y}_k|g=i)p(g=i)}$

end for

for $i,j=1\rightarrow G$ **do** {Likelihood Ratios}

$L_{i,j}\leftarrow\frac{\prod_{l=1}^{k}p(\mathbf{y}_{(l)}|g=i)}{\prod_{l=1}^{k}p(\mathbf{y}_{(l)}|g=j)}\frac{p(g=i)}{p(g=j)}$

end for

if $\forall_{i\neq j}\ L_{i,j}>\eta$ **then** {Classification}

$\gamma\leftarrow i$

return $\gamma,\mathbf{\Phi}_k$

end if

end while

where $p(1,\ldots,g-1,g+1,\ldots,G)$ represents the misclassification error probability, when g is the true Gaussian. This is precisely what SHT provides, based on Bayesian updates of the prior class probabilities and maximum likelihood. Algorithm 1 describes in detail the SHT method [28], [41].

We have provided an algorithm, based on hypothesis testing, to automatically determine the number of measurements required in the first step of the two-step framework, for a given probability of classification error. This in conjunction with AIDA, provides the AIDA-SHT CS protocol cited in Table I.

V. Experimental Results

A. Non-Adaptive Statistical Compressive Sensing

We compare the reconstruction performance of the proposed non-adaptive sensing RIP-AB (Section III), with random sensing, the optimal non-adaptive sensing for unstructured dictionaries [9], [11], and the optimal sensing for structured dictionaries [10]. We use all 8×8 overlapping patches from 20 natural images taken from the Berkeley segmentation dataset [42] in order to (offline) adapt the learned dictionaries (GMM) and sensing matrices (which depend on the PCA basis of the GMM, except for

random sensing of course) to each image. The (online) evaluation of the learned dictionaries and non-random sensing matrices was done using non-overlapping 8×8 patches. Only κ = 11 iterations of the MAP-EM statistical CS were needed to learn the dictionaries and non-random sensing matrices.[3]

TABLE II
MEAN IMAGE RECONSTRUCTION PSNR (DBS), NON-OVERLAPPING PATCHES.

Compression Ratio	**Sensing Matrix**			
	Random	**Optimal Unstructured**	**Optimal Structured**	**RIP-AB GMM**
8.0	28.04	28.74	28.93	29.30
5.3	29.86	30.42	30.59	30.98
3.2	32.90	33.47	33.53	34.0

Table II shows the mean peak signal to noise ratio (PSNR) of the reconstructed images (from the reconstructed non-overlapping patches), for three levels of signal compression, where $PSNR = 10\ \log_{10}\left(\frac{I^2_{max}}{MSE}\right)$, I_{max} corresponding to the maximum image intensity. The RIP-AB sensing strategy achieves the best reconstruction PSNRs, as expected. Using a paired t-test, we found that the differences are statistically significant, with a p-value less than 0.001. Figure 1 shows the original and reconstructed 8×8 non-overlapping patches for a selected image (results with other selected images are also shown in the supplementary material, figures S1-S2). The quantitative improvements are visually observed as well.

B. Adaptive Statistical Compressive Sensing

We now compare the classification and reconstruction performance of the proposed two-step adaptive statistical CS framework (Section IV), with several configurations as indicated in Table I. For this purpose, we generate two-class synthetic Gaussian signals of dimensions 36, 64, and 100, where the Bhattacharyya distance (BD) between them is varied as well, see Appendix E for details.

In addition, we also use 6×6, 8×8, and 10×10 non-overlapping patches extracted from 50 natural images from the Berkeley segmentation dataset [42]. Three levels of noise were also considered: no added noise, noise level of 40 dbs, and noise level of 30 dbs, for both the synthetic and natural image signals. There are 19 learned classes in this case, used to efficiently model patches of natural images (see [13] for details).

[3]We require more MAP-EM iterations than those indicated on Section II, since now, the sensing matrix $\mathbf{\Phi}$ and the dictionary (GMM) are both being simultaneously adapted.

The dictionary is here not adapted to each image, in order to provide the same GMM to all the considered two-step configurations (Table I), so that the differences between configurations are due only to the sensing itself and not due to the adapted dictionaries.

1) Synthetic Signals: Figure 2 shows the classification error at Step 1 and the corresponding MSE at Step 2, for the indicated BDs and three levels of noise. As expected, the best classification results are obtained for IDA and AIDA-SHT, being AIDA-SHT the best, stopping automatically between two and three samples, on average. The worst classification results were obtained for the RIP-AB non-adaptive sensing matrix and random sensing. It can be noticed that despite the relatively bad classification performance of RIP-AB, the reconstruction is equal or better than random and it improves as $K \to M$, as expected, since RIP-AB is better than random for non-adaptive single-step batch CS (sections III and V-A), i.e., when $K = M$. RIP-AB does better than random on the final reconstruction, using $K < M$ measurements, because the BDs between the two Gaussians is relatively small, hence, selecting the wrong Gaussian will not be that harmful for reconstruction during the second step. The lowest reconstruction errors are achieved using AIDA-SHT in the first step and the optimal (in the MI sense) adaptive sensing in the second step. However, as the noise increases, the reconstruction performance of AIDA-SHT degrades with respect to IDA. The reason for this is that IDA uses only the (clean) information provided by the covariance matrices, while SHT tests hypotheses based on the noisy CS signals. Nevertheless, AIDA-SHT stops automatically, without knowing a priori the number of steps K required in the first step, while IDA (or AIDA alone) requires a pre-defined number of steps K.

It can also be observed in Figure 2 that using IDA in the first step and the non-adaptive optimal for the selected Gaussian in the second step is worse than using random in the first step and the non-adaptive optimal for the selected Gaussian in the second, despite the fact that IDA has a very good classification performance in the first step. The reason for this is that IDA uses information from the average covariance matrix, which is likely to lead to a sensing matrix in the first step that is correlated with the sensing matrix in the second, especially when the number of classes is small (two in this case). Hence, the importance of adaptivity in the second step. This does not affect random sensing, since it is very unlikely that a random sensing matrix will be correlated with the non-adaptive optimal sensing matrix in the second step. This does not seem to affect the RIP-AB sensing, probably because it does not use the average covariance matrix either.

Figure 3 shows the classification error at the first step and the corresponding reconstruction MSE at the second step, for larger BDs than those in Figure 2, with the same noise levels. All classification accuracies are now better than in Figure 2, since now the distance between the covariance matrices is larger, making

classification easier. AIDA-SHT still has the best classification accuracy, but the difference with IDA is reduced. In the second step, the reconstruction performance using AIDA-SHT in the first step deteriorates more with noise, and in this case IDA is better. We believe the reason for this is that the adaptive sensing matrix of AIDA is much better than IDA for classification, but good features for classification are not necessarily good for reconstruction. In addition, since at larger BD distances classification becomes easier, the need for a good classifier in the first step reduces, while reconstruction becomes more important.

Due to space limitations, the results for other signal sizes and BDs are only shown in the supplementary material. However, and as can be seen from the figures in the supplementary material (figures S3-S40), the behavior is quite similar to the results just presented.

2) Patches from Natural Images: Figure 4 shows the mean PSNR for the non-overlapping 6×6 patches extracted from 50 natural images at the three noise levels considered. Here we do not know the right class for each one of the signals, hence, we only report the reconstruction accuracy in the second step. The best reconstruction performance in the noise-free case is obtained for AIDA-SHT, and the classification (first step) stops between 5 to 6 samples on average. As the noise increases, the difference between AIDA and IDA reduces. This follows the same behavior observed with synthetic data, where AIDA-SHT degrades with noise. There is a wave-like behavior on the non-adaptive protocols used, with the exception of random sensing, probably due to the possible coherence of the sensing matrix between steps, which depends on K. Results for non-overlapping patches of sizes 8×8 and 10×10 can be found in the supplementary material (figures S41 and S42).

Figure 5 shows the original and reconstructed 6×6 non-overlapping patches for a selected image using the different two-step CS configurations considered (Table I), $K = 5$, and no noise added (AIDA-SHT uses $K = 5.41$ on average). Using IDA and AIDA-SHT on the first step and the optimal adaptive (MI) on the second step produces the best reconstruction performances, with PSNRs that are ~ 6 dbs above random sensing in the first step. Notice that the reported reconstructions do not use dictionary (and sensing) adaptation, explaining the relative bad quality of the reconstructed patches using random and RIP-AB sensing in the first step (compare to the results reported on Section V-A). Results with other selected images and patch sizes can be found in the supplementary materials (figures S43-S53).

Note that the results using the proposed two-step framework (figures 2-4 and S3-S42) also include the single-step batch results for random, RIP-AB, and IDA, for the particular case when $K = M$. Hence, from these figures, we can conclude that in general the proposed two-step framework does provide better reconstructions than single-step (batch) random, RIP-AB, and IDA. AIDA-SHT should not be done in batch mode, since it is fully adaptive.

In summary, AIDA-SHT in the first step and the adaptive optimal (MI) sensing matrix in the second is the best sensing protocol for the two-step SCS proposed, provided that the noise level is not too high. A good alternative (less expensive computationally) to AIDA-SHT in the first step is the well-known non-adaptive batch IDA, provided that we know a priori the optimal number of measurements K needed for class detection in the first step.

C. Classification

In this section we test on the Statlog (Landsat Satellite) dataset, consisting of six classes with 36 numerical attributes and 6435 instances [43]. For this dataset, we used dictionary learning and data-adaptation. As Figure 6a shows, only IDA achieves good classification accuracy when training the dictionaries (GMM with 6 classes). Hence, in order to remove the effect of dictionary adaptation, we used the best learned dictionary for this dataset (obtained using IDA) for all the remaining sensing matrices. Figure 6b shows the classification accuracy using this GMM, IDA and AIDA-SHT achieve the best classification performance. AIDA-SHT, automatically stops at five measurements on average.

These classification results correspond to the classification of each CS signal independently of the other CS signals, and within the proposed two-step framework. Therefore, it cannot be directly compared with classifiers that use all the CS data at the same time.[4] These results show that classification accuracy closely follows the results obtained using synthetic data, but now, for real data. In addition, and if we are only interested in sensing a given class, we can stop sensing that signal if we determine in the first step that the class is of no interest, reducing thus the average number of measurements. Indeed, the total number of measurements for single step (adaptive or not) sensing, with S signals and M measurements, is SM, but, if we are only interested in a given class γ, the average number of measurements reduces to $S(K(1-p(\gamma))+Mp(\gamma))$, where $p(\gamma)$ is the probability of class γ.

Finally, the proposed AIDA or AIDA-SHT could be used for classification purposes only (single step), where the final classification is achieved using an offline classifier that uses all the CS data, such as support vector machines, neural networks, or Bayesian classifiers.

VI. Conclusion

We have developed a novel two-step SCS framework tailored to GMMs, from where several SCS protocols can be chosen (Table I). The best sensing protocol, in terms of accuracy of classification in

[4]If we take, however, all the CS signals and use a quadratic Bayesian classifier, we obtain a classification accuracy of 82% that is similar to the results reported in [37].

the first step, reconstruction in the second step, and automatic selection of number of samples, is the AIDA-SHT in the first step and the optimal (MI) adaptive sensing in the second step. A good alternative to AIDA-SHT in the first step is a batch non-adaptive IDA or a batch AIDA with a predefined number of measurements, which seems more robust to noise and deviations from the GMM assumption. The two-step SCS is clearly superior to a batch single step SCS, the reconstruction accuracy is equal or better than a single step SCS and the average total number of measurements is lower, we can stop sensing a given signal if in the first step we determine that the signal is of no interest.

The two-step SCS framework presented here also applies to a single step adaptive or batch sensing approach, for the particular case when $K = M$. More general GMMs than the single Gaussian per signal studied here, or other non-Gaussian models, can also be considered within this framework, at the cost of higher mathematical and possibly computational complexity.

APPENDIX A
SOLUTION TO EQUATION (14)

The general orthogonal Procrustes problem is defined as [36]

$$\mathbf{X} = \text{argmin}_{\hat{\mathbf{X}}} \|\mathbf{A}\hat{\mathbf{X}} - \mathbf{C}\|_F^2, \ s.t. \ \hat{\mathbf{X}}\hat{\mathbf{X}}^T = \mathbf{I}. \tag{36}$$

Let $\mathbf{M} = \mathbf{A}^T\mathbf{C}$, and $\mathbf{M} = \mathbf{U}\boldsymbol{\Delta}\mathbf{W}^T$ its eigen-decomposition. The solution to (36) is given by $\mathbf{X} = \mathbf{U}\mathbf{W}^T$. Now, noticing that $\|\mathbf{A}\|_F^2 = \|\mathbf{A}^T\|_F^2$, we can rewrite (14) as

$$\mathbf{B}^T = \text{argmin}_{\hat{\mathbf{B}}} \left\{ \left\| [\mathbf{E}^T\hat{\mathbf{B}}^T - \mathbf{I}_N \right\|_F^2 \right\} \ s.t. \ \hat{\mathbf{B}}^{\mathbf{T}}\hat{\mathbf{B}} = \mathbf{I}_N. \tag{37}$$

Comparing (37) and (36) one can see that $\mathbf{M} = \mathbf{E}$, and if $\mathbf{E} = \mathbf{U}\boldsymbol{\Delta}\mathbf{W}^T$, then the solution to (37) is $\mathbf{B}^T = \mathbf{U}\mathbf{W}^T$, hence, $\mathbf{B} = \mathbf{W}\mathbf{U}^T$.

APPENDIX B
DERIVATION OF EQUATION (18)

By Bayes' theorem, we know that

$$p(\mathbf{y}_k|g, \mathbf{y}_{(k-1)}) = \frac{p(\mathbf{y}_k, g|\mathbf{y}_{(k-1)})}{p(g|\mathbf{y}_{(k-1)})}. \tag{38}$$

Hence, (17) becomes

$$I(\mathbf{y}_k; g|\mathbf{y}_{(k-1)}) = E\Big\{ \log \frac{p(\mathbf{y}_k|g, \mathbf{y}_{(k-1)})}{p(\mathbf{y}_k|\mathbf{y}_{(k-1)})} \Big\}. \tag{39}$$

Also by Bayes,

$$p(\mathbf{y}_k|g,\mathbf{y}_{(k-1)}) = \frac{p(\mathbf{y}_k,\mathbf{y}_{(k-1)}|g)}{p(\mathbf{y}_{(k-1)}|g)} = \frac{p(\mathbf{y}_{(k)}|g)}{p(\mathbf{y}_{(k-1)}|g)}, \tag{40}$$

$$p(\mathbf{y}_k|\mathbf{y}_{(k-1)}) = \frac{p(\mathbf{y}_k,\mathbf{y}_{(k-1)})}{p(\mathbf{y}_{(k-1)})} = \frac{p(\mathbf{y}_{(k)})}{p(\mathbf{y}_{(k-1)})}. \tag{41}$$

Replacing (40) and (41) in (39), and applying logarithm properties,

$$\begin{aligned} I(\mathbf{y}_k;g|\mathbf{y}_{(k-1)}) &= E\{\log\ p(\mathbf{y}_{(k)}|g)\} - E\{\log\ p(\mathbf{y}_{(k)})\} \\ &\quad -\big(E\{\log\ p(\mathbf{y}_{(k-1)|g})\} - E\{\log\ p(\mathbf{y}_{(k-1)})\}\big). \end{aligned} \tag{42}$$

Applying the definition of differential entropy, $H(f) = -E\{\log\ f\}$, we arrive to Equation (18).

APPENDIX C
DERIVATION OF EQUATION (27)

The covariance of $\mathbf{y}_{(k)}$, given Gaussian class g, is given by

$$|\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g}| = |\mathbf{\Phi}_{(k)}\mathbf{\Sigma}_g\mathbf{\Phi}_{(k)}^T + \sigma^2\mathbf{I}_{kb}| = \left|\begin{bmatrix}\mathbf{\Phi}_{(k-1)}\\ \mathbf{\Phi}_k\end{bmatrix}\mathbf{\Sigma}_g\begin{bmatrix}\mathbf{\Phi}_{(k-1)}^T & \mathbf{\Phi}_k^T\end{bmatrix} + \sigma^2\mathbf{I}_{kb}\right|$$

$$= \begin{vmatrix}\mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g\mathbf{\Phi}_{(k-1)}^T + \sigma^2\mathbf{I}_{(k-1)b} & \mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g\mathbf{\Phi}_k^T \\ \mathbf{\Phi}_k\mathbf{\Sigma}_g\mathbf{\Phi}_{(k-1)}^T & \mathbf{\Phi}_k\mathbf{\Sigma}_g\mathbf{\Phi}_k^T + \sigma^2\mathbf{I}_b\end{vmatrix} = \begin{vmatrix}\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g} & \mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g\mathbf{\Phi}_k^T \\ \mathbf{\Phi}_k\mathbf{\Sigma}_g\mathbf{\Phi}_{(k-1)}^T & \mathbf{\Phi}_k\mathbf{\Sigma}_g\mathbf{\Phi}_k^T + \sigma^2\mathbf{I}_b\end{vmatrix}.$$

Since by hypothesis $\mathbf{x}$ is an N-dimensional Gaussian with positive definite covariance $\mathbf{\Sigma}_g$ (g unknown of course), then the covariance matrix $\mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g\mathbf{\Phi}_{(k-1)}^T$ is also positive definite [38], hence, invertible and we can use the identity for determinants $\begin{vmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{vmatrix} = |\mathbf{A}||\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}|$ [40], [44]. Then,

$$\begin{aligned} |\mathbf{\Sigma}_{\mathbf{y}_{(k)}|g}| &= |\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g}|\ |\mathbf{\Phi}_k\mathbf{\Sigma}_g\mathbf{\Phi}_k^T + \sigma^2\mathbf{I}_b - \mathbf{\Phi}_k\mathbf{\Sigma}_g(\mathbf{\Phi}_{(k-1)})^T\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g}^{-1}\mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g\mathbf{\Phi}_k^T| \\ &= |\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g}|\ \left|\mathbf{\Phi}_k\big(\mathbf{\Sigma}_g - \mathbf{\Sigma}_g\mathbf{\Phi}_{(k-1)}^T\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}|g}^{-1}\mathbf{\Phi}_{(k-1)}\mathbf{\Sigma}_g + \sigma^2\mathbf{I}_b\big)\mathbf{\Phi}_k^T\right|, \end{aligned} \tag{43}$$

where in the last step we used the fact that $\mathbf{I}_b = \mathbf{\Phi}_k\mathbf{\Phi}_k^T$.

Starting from (24), and following the same steps indicated before, it is straightforward to show that

$$|\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k)}}| = |\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k-1)}}|\left|\mathbf{\Phi}_k\big(\bar{\mathbf{\Sigma}} - \bar{\mathbf{\Sigma}}\mathbf{\Phi}_{(k-1)}^T\bar{\mathbf{\Sigma}}_{\mathbf{y}_{(k-1)}}^{-1}\mathbf{\Phi}_{(k-1)}\bar{\mathbf{\Sigma}} + \sigma^2\mathbf{I}_b\big)\mathbf{\Phi}_k^T\right|. \tag{44}$$

Replacing (43) and (44) in (25), and defining $\mathbf{P}_g$ and $\bar{\mathbf{P}}$ as indicated on (27), we arrive at Equation (27).

APPENDIX D
DERIVATION OF EQUATION (34)

By Bayes' theorem,

$$\frac{p(\mathbf{y}_k^\gamma, \mathbf{x}^\gamma | \mathbf{y}_{(k-1)}^\gamma)}{p(\mathbf{x}^\gamma | \mathbf{y}_{(k-1)}^\gamma)} = p(\mathbf{y}_k^\gamma | \mathbf{x}^\gamma, \mathbf{y}_{(k-1)}^\gamma) = \frac{p(\mathbf{y}_k^\gamma, \mathbf{y}_{(k-1)}^\gamma | \mathbf{x}^\gamma)}{p(\mathbf{y}_{(k-1)}^\gamma | \mathbf{x}^\gamma)} = \frac{p(\mathbf{y}_{(k)}^\gamma | \mathbf{x}^\gamma)}{p(\mathbf{y}_{(k-1)}^\gamma | \mathbf{x}^\gamma)}. \tag{45}$$

Also, by Bayes,

$$p(\mathbf{y}_k^\gamma | \mathbf{y}_{(k-1)}^\gamma) = \frac{p(\mathbf{y}_k^\gamma, \mathbf{y}_{(k-1)}^\gamma)}{p(\mathbf{y}_{(k-1)}^\gamma)} = \frac{p(\mathbf{y}_{(k)}^\gamma)}{p(\mathbf{y}_{(k-1)}^\gamma)}. \tag{46}$$

Replacing (45) and (46) in (34),

$$I(\mathbf{y}_k^\gamma; \mathbf{x}^\gamma | \mathbf{y}_{(k-1)}^\gamma) = E\Big\{ \log \frac{p(\mathbf{y}_{(k)}^\gamma | \mathbf{x}^\gamma) p(\mathbf{y}_{(k-1)}^\gamma)}{p(\mathbf{y}_{(k-1)}^\gamma | \mathbf{x}^\gamma) p(\mathbf{y}_{(k)}^\gamma)} \Big\}, \tag{47}$$

and applying logarithm properties and the definition of differential entropy,

$$I(\mathbf{y}_k^\gamma; \mathbf{x}^\gamma | \mathbf{y}_{(k-1)}^\gamma) = H\big(p(\mathbf{y}_{(k)}^\gamma)\big) - H\big(p(\mathbf{y}_{(k)}^\gamma | \mathbf{x}^\gamma)\big) - \Big[H\big(p(\mathbf{y}_{(k-1)}^\gamma)\big) - H\big(p(\mathbf{y}_{(k-1)}^\gamma | \mathbf{x}^\gamma)\big)\Big]. \tag{48}$$

The term within brackets in (48) is independent of $\mathbf{\Phi}_k$, so it can be considered a constant. The probability $p(\mathbf{y}_{(k)}^\gamma) = p(\mathbf{y}_{(k)} | g = \gamma)$ is the same as (19), replacing g by γ, and its entropy is given by [35]

$$H\big(p(\mathbf{y}_{(k)}^\gamma)\big) = \frac{1}{2}\Big[kb(1 + \log(2\pi)) + \log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma}|\Big], \ \mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma} = \mathbf{\Phi}_{(k)} \mathbf{\Sigma}_\gamma \mathbf{\Phi}_{(k)}^T. \tag{49}$$

On the other hand,

$$p(\mathbf{y}_{(k)}^\gamma | \mathbf{x}^\gamma) = \frac{1}{(2\pi\sigma^2)^{kb/2}} exp\Big(- \frac{1}{2\sigma^2} (\mathbf{y}_{(k)}^\gamma - \mathbf{\Phi}_{(k)} \mathbf{x}^\gamma)^T (\mathbf{y}_{(k)}^\gamma - \mathbf{\Phi}_{(k)} \mathbf{x}^\gamma) \Big), \tag{50}$$

which has entropy $H(p(\mathbf{y}_{(k)}^\gamma | \mathbf{x}^\gamma)) = \frac{kb}{2} \log(2\pi\sigma^2)$, and is independent of $\mathbf{\Phi}_k$. Hence, Equation (33) can be rewritten as,

$$I(\mathbf{y}_k^\gamma; \mathbf{x}^\gamma | \mathbf{y}_{(k-1)}^\gamma) = \frac{1}{2} \log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma}| + F(\mathbf{y}_{(k-1)}^\gamma), \tag{51}$$

where $F(\mathbf{y}_{(k-1)}^\gamma)$ accounts for all terms dependent on $\mathbf{y}_{(k-1)}^\gamma$ plus some constants. Hence, in order to solve (32), we need to maximize $\log|\mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma}|$. Since the logarithm is a monotonic function, it is enough to maximize $|\mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma}|$. Now, following the same steps indicated in (43), we arrive at

$$|\mathbf{\Sigma}_{\mathbf{y}_{(k)}^\gamma}| = |\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}^\gamma}| \ \Big| \mathbf{\Phi}_k \big(\mathbf{\Sigma}_\gamma - \mathbf{\Sigma}_\gamma \mathbf{\Phi}_{(k-1)}^T \mathbf{\Sigma}_{\mathbf{y}_{(k-1)}^\gamma}^{-1} \mathbf{\Phi}_{(k-1)} \mathbf{\Sigma}_\gamma + \sigma^2 \mathbf{I}_b \big) \mathbf{\Phi}_k^T \Big|. \tag{52}$$

Let $\mathbf{P}_\gamma = \mathbf{\Sigma}_\gamma - \mathbf{\Sigma}_\gamma \mathbf{\Phi}_{(k-1)}^T \mathbf{\Sigma}_{\mathbf{y}_{(k-1)}^\gamma}^{-1} \mathbf{\Phi}_{(k-1)} \mathbf{\Sigma}_\gamma + \sigma^2 \mathbf{I}_b$, and since $|\mathbf{\Sigma}_{\mathbf{y}_{(k-1)}^\gamma}|$ is independent of $\mathbf{\Phi}_k$, the solution to (32), reduces to,

$$\mathbf{\Phi}_k = \text{argmax}_{\hat{\mathbf{\Phi}}_k} \Big| \hat{\mathbf{\Phi}}_k \mathbf{P}_\gamma \hat{\mathbf{\Phi}}_k^T \Big|, \ s.t. \ \hat{\mathbf{\Phi}}_k \hat{\mathbf{\Phi}}_k^T = \mathbf{I}. \tag{53}$$

This corresponds to a high dimensional extension of the Rayleigh-Ritz Theorem, whose solution is given by Equation (34) (see proof of this theorem in [44]).

APPENDIX E
SYNTHETIC MULTIDIMENSIONAL GAUSSIAN DISTRIBUTIONS

Let $\mathbf{R}$ be an $N \times N$ Gaussian random matrix and $\mathbf{R} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T$ its corresponding singular value decomposition. We define synthetic covariance matrices as $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, $\mathbf{\Lambda}_{ii} = r10^{\beta} i^{-\omega}$, where $r \in (0\ 1], \beta \in [48], \omega \in \{3, 4\}$. These ranges where chosen empirically, in order to obtain Bhattacharyya distances (BDs) in the range $[30\ +\infty]$, where the BD is given by $BD = \frac{1}{2}\ln(\frac{|\Sigma_0|}{\sqrt{|\Sigma_0||\Sigma_1|}})$.

More specifically, we obtained around 100 pairs of covariances matrices with BDs for each one of the following ranges: [30 46), [46 62), [62 78), [78 94), [94 110], (110 126], (126 142], and $(142\ +\infty)$, for a total of 800 pairs of covariance matrices. Bhattacharyya distances among patches from natural images are in the [30 61] range, which justifies the chosen ranges. Larger Bhattacharyya distances than those considered here are quite difficult to obtain at random and often lead to numerical instability.

Acknowledgments: Work supported by ONR, NSF, DARPA, NGA, ARO, and NSSEFF. We thank very constructive comments and discussions with Prof. Robert Calderbank, Prof. David Brady, and Prof. Stanley Osher.

REFERENCES

[1] E. J. Candes and M. B. Wakin, "People hearing without listening: An introducton to compressive sampling," *IEEE Signal Proc. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[2] G. Peyré, "Best basis compressed sensing," *IEEE Trans. Image Proc.*, vol. 58, no. 5, pp. 2613–2622, 2010.

[3] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[4] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 2, no. 5, pp. 489–509, Feb. 2006.

[5] E. J. Candes, "Compressive sampling," in *Proc. Int. Congr. Math.*, Madrid, Spain, 2006, pp. 1433–1452.

[6] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, no. 4-6, pp. 918–925, 2007.

[7] S. D. Howard, R. Calderbank, and S. J. Searle, "A fast reconstruction algorithm for deterministic compressive sensing using second order Redd-Muller codes," in *42nd Annual Conference on Information Sciences and Systems*, IEEE, Ed., Princeton, NJ, March 2008, pp. 11–15.

[8] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Proc.*, vol. 55, no. 12, pp. 5695–5702, 2007.

[9] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Proc.*, vol. 18, no. 7, pp. 1395–1408, 2009.

[10] K. Rosenblum, L. Zelnik-Manor, and Y. C. Eldar. (2010, Sep.) Sensing matrix optimization for block-sparse decoding. arXiv:1009.1533v1.

[11] J. M. Duarte-Carvajalino, G. Yu, L. Carin, and G. Sapiro, "Adapted statistical compressive sensing: Learning to sense Gaussian mixture models," ICASSP 2012, to appear.

[12] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Proc.*, vol. 17, no. 1, pp. 53–69, 2008.

[13] G. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture modesl to structured sparsity," *IEEE Trans. Image Proc.*, to appear, 2011.

[14] M. F. Duarte and Y. C. Eldar. (2011, Jun.) Structured compressed sensing: from theory to applications. arXiv:1106.6224v2.

[15] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Proc.*, vol. 56, no. 6, pp. 2334–2345, 2008.

[16] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1872–1882, April 2009.

[17] D. Needell and R. Vershynin. (2007, Jul.) Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. arXiv:0707.4203v4.

[18] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[19] G. Yu and G. Sapiro, "Statistical compressed sensing of Gaussian mixture models," *IEEE Trans. Signal Proc.*, vol. 59, no. 12, pp. 5842–5858, 2011.

[20] M. Chen, J. Silva, C. Paisley, D. D. Wang, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Proc.*, vol. 58, no. 12, pp. 6140–6155, 2010.

[21] A. Averbuch, S. Dekel, and S. Deutsch, "Adaptive compressed image sensing using dictionaries," to appear on SIAM J. Imaging Sciences.

[22] J. Haupt, R. Nowak, and R. M. Castro, "Adaptive sensing for sparse recovery," in *DSP/SPE*, Marco Island, FL, 2009, pp. 702–707.

[23] J. Haupt, R. M. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.

[24] A. Aldroubi, H. Wang, and K. Zarringhalam. (2008, Oct.) Sequential adaptive compressed sampling via Huffman codes. arXiv:0810.4916v2.

[25] A. Ashok, P. K. Baheti, and M. A. Neifeld, "Compressive imaging system design using task-specific information," *Applied Optics*, vol. 47, no. 25, pp. 4457–4471, 2008.

[26] J. Ke, A. Ashok, and M. A. Neifeld, "Object reconstruction from adaptive compressive measurements in feature-specific imaging," *Applied Optics*, vol. 49, no. 34, pp. H27–H39, 2010.

[27] W. R. Carson, M. R. D. Rodrigues, M. Chen, L. Carin, and R. Calderbank, "How to focus the discriminative power of a dictionary," ICASSP 2012, to appear.

[28] P. K. Baheti and M. A. Neifeld, "Recognition using information-optimal adaptive feature-specific imaging," *J. Opt. Soc. Am. A*, vol. 26, no. 4, pp. 1055–1070, 2009.

[29] R. Calderbank, L. Carin, and M. Chen, In preparation, private communication, 2011.

[30] A. Wald, "Sequential analysis of statistical hypotheses," *Math. Stat.*, vol. 16, no. 2, pp. 117–176, 1945.

[31] P. Armitage, "Sequential analysis with more than two alternative hypotheses and its relation to discriminant function," *J. R. Stat. Soc.*, vol. 12, no. 1, pp. 137–144, 1950.

[32] D. Guo, S. Shlomo, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[33] D. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, 2006.

[34] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., ser. Telecommunications, D. L. Schilling, Ed. New York, NY: John Wiley and Sons, 1991.

[36] P. H. Schonemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, March 1966.

[37] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1394–1407, 2007.

[38] H. Stark and W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*. Prentice-Hall, 1994.

[39] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 512–519, 2005.

[40] K. B. Petersen and M. S. Pedersen. (2008, Oct.) The matrix cookbook. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[41] P. K. Baheti and M. A. Neifeld, "Adaptive feature-specific imaging: a face recognition example," *Applied Optics*, vol. 47, no. 10, pp. B21–B31, 2008.

[42] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *8th Int. Conf. Computer Vision*, vol. 2. Vancouver, BC, Canada: IEEE, 2001, pp. 416–423.

[43] A. Frank and A. Asuncion. (2010) UCI machine learning repository. http://archive.ics.uci.edu/ml.

[44] M. Brookes. (2011, Dec.) The matrix reference manual. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html

(a) (b)

(c) (d)

Fig. 1. Reconstructed image from learned dictionaries and non-overlapping patches of size 8×8 (CS to 12 samples). a) Original, b) Random (29.1 dbs), c) Unstructured/Structured (29.2 dbs), d) RIP-AB (32.1 dbs)

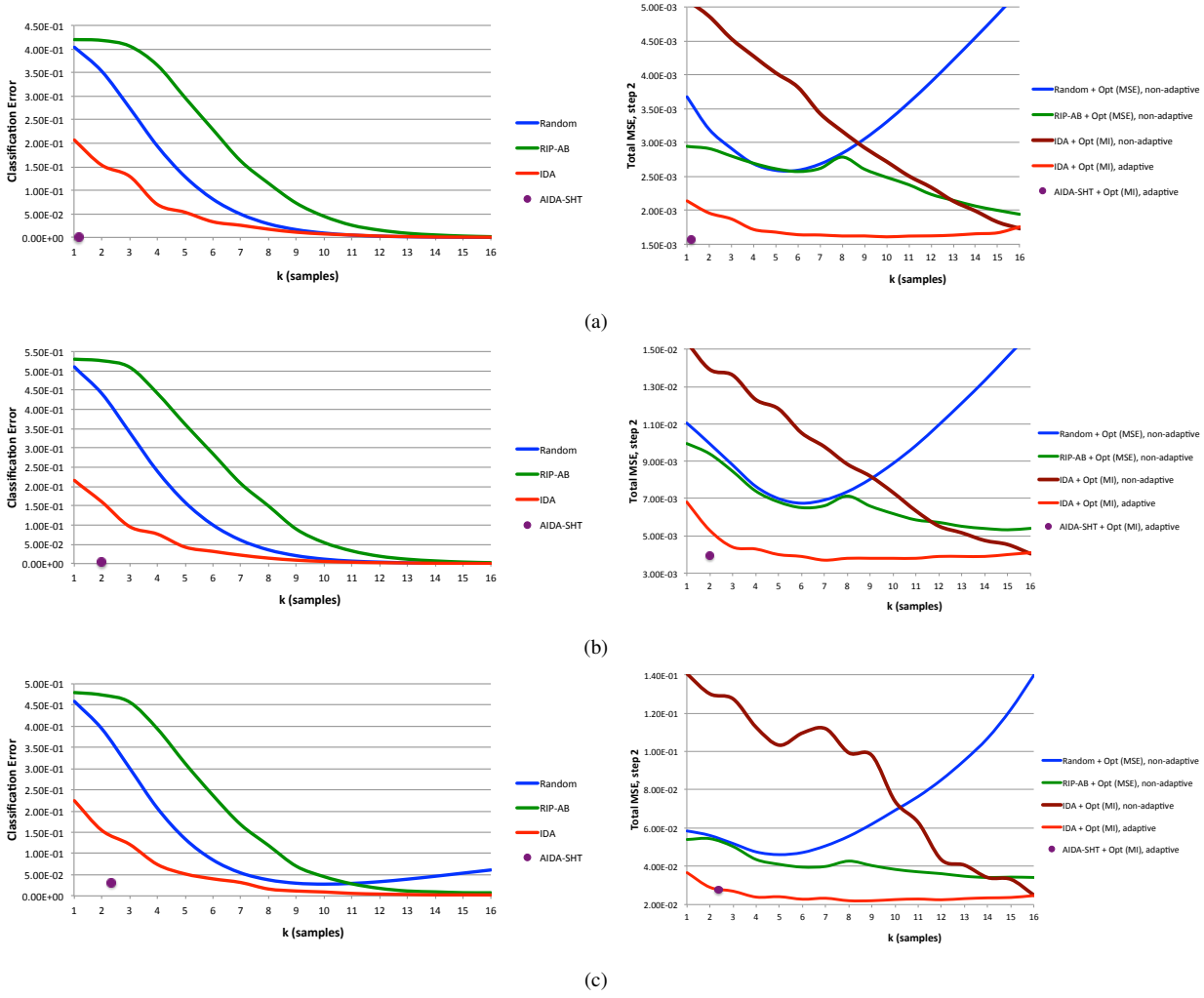
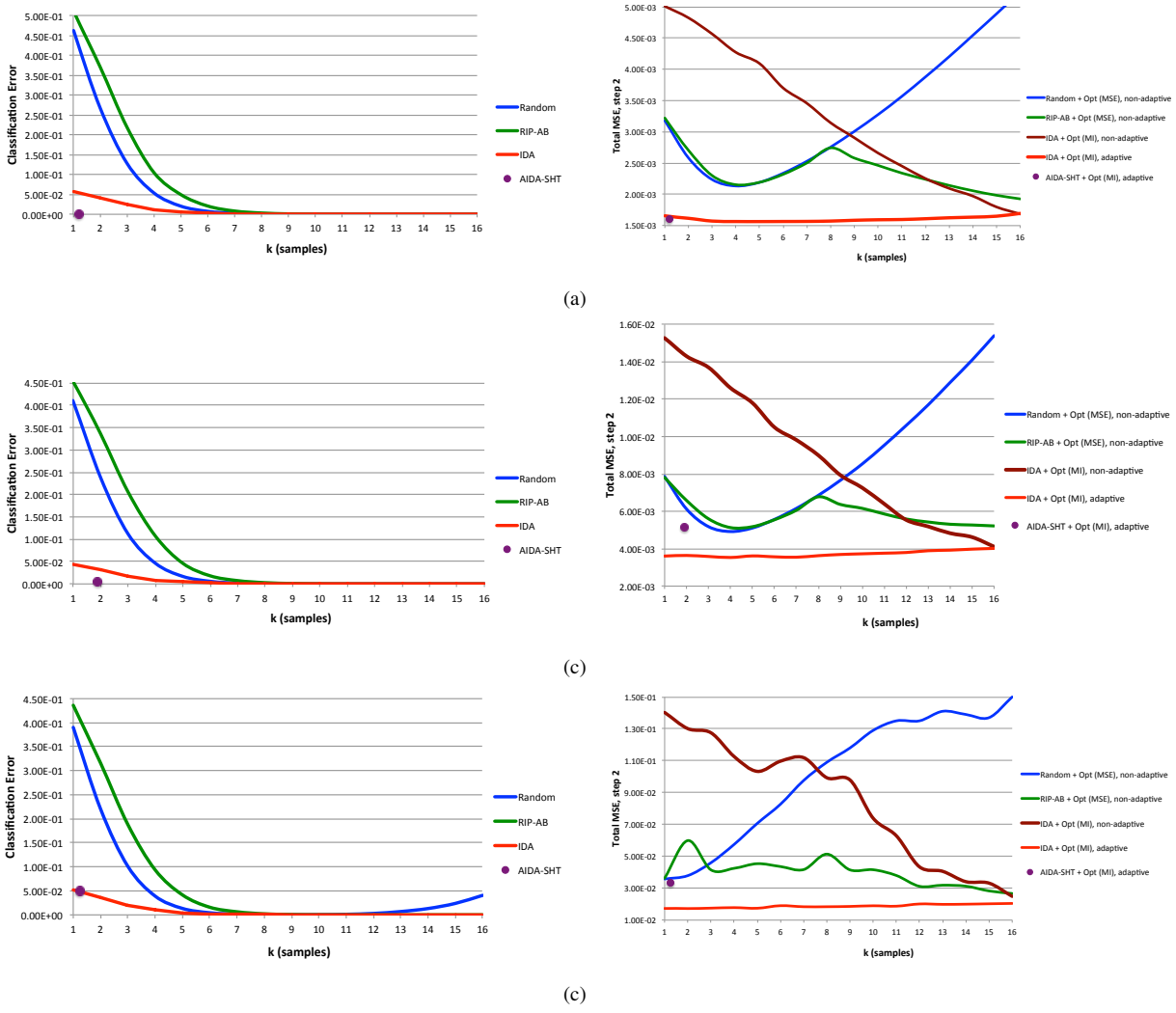


Fig. 2. Classification accuracy (Step 1) and reconstruction MSE (step 2) for synthetic signals of dimension 64 (CS to 16 samples) and BDs $\in [30\ 46)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.

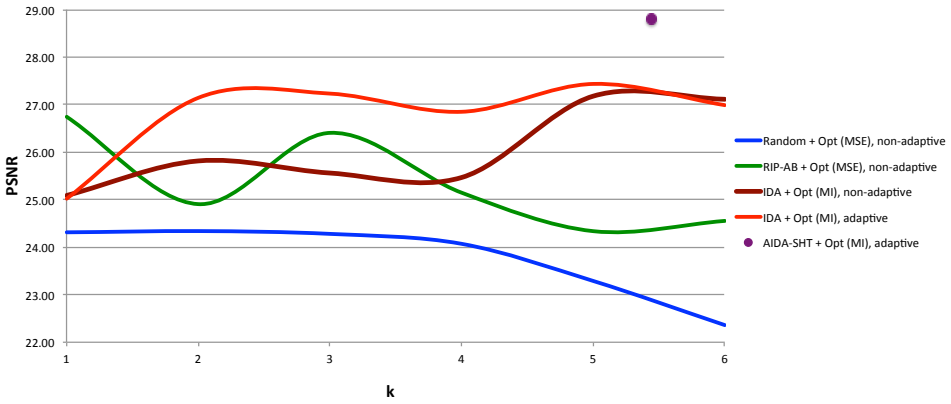


(a)

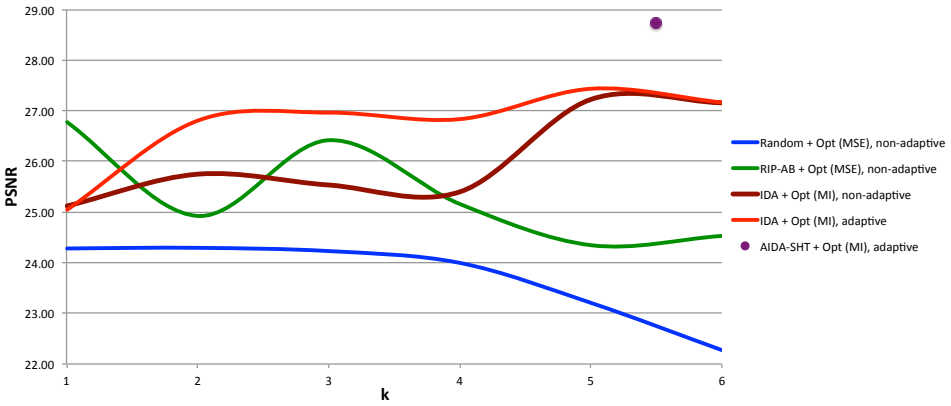
(c)

(c)

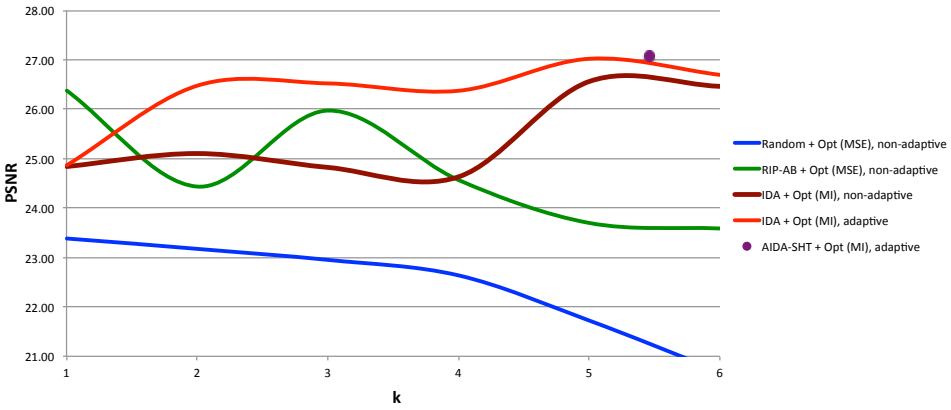
Fig. 3. Classification accuracy (Step 1) and reconstruction MSE (step 2) for synthetic signals of dimension 64 (CS to 16 samples) and BDs $\in [62\ 78)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)



(b)



(c)

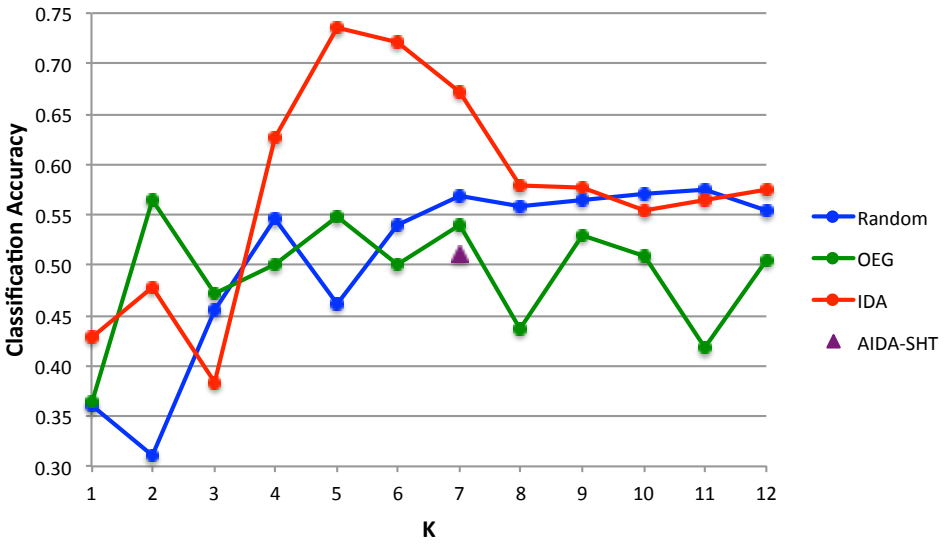
Fig. 4. PSNR (Step 2) of reconstructed natural images with non-overlapping patches of size 6×6 (CS to 6 samples). a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.

(a) (b)

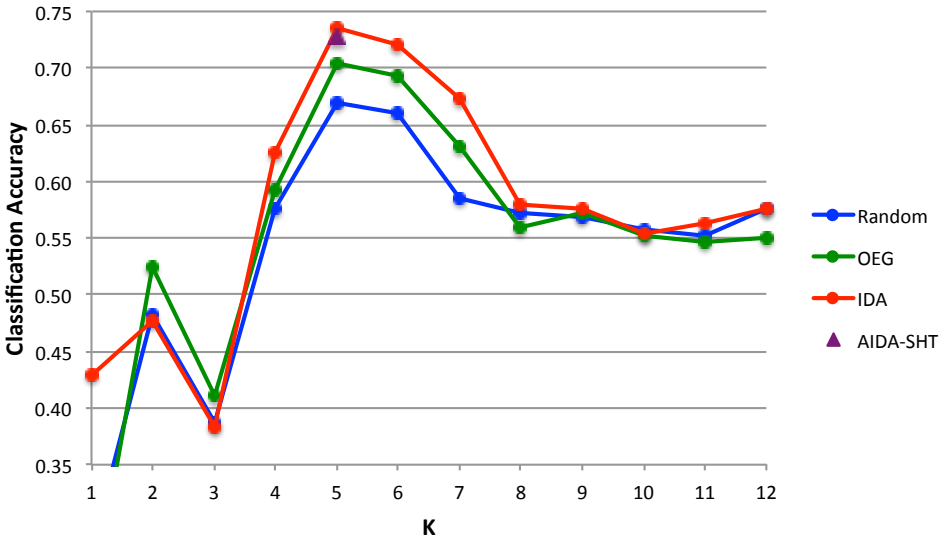
(c) (d)

(e) (f)

Fig. 5. Reconstructed image from non-overlapping patches of size 6×6 (CS to 6 samples) using the following two-step protocols: a) Original, b) Random - Optimum (MSE) non-adaptive (26.8 dbs), c) RIP-AB - Optimum (MSE) non-adaptive (27.66 dbs), d) IDA - Optimum (MSE) non-adaptive (30.51 dbs), e) IDA- Optimum (MI) adaptive (30.77 dbs), and f) AIDA-SHT-Optimum (MI) adaptive (33.9 dbs).



(a)



(b)

Fig. 6. Classification accuracy for the satellite dataset (CS to 6 samples) using a) dictionary learning, b) best learned dictionary.

Supplementary Material

Task-Driven Adaptive Statistical Compressive Sensing of Gaussian Mixture Models

Julio M. Duarte-Carvajalino, Guoshen Yu, Lawrence Carin, and Guillermo Sapiro

Computational Complexity

We analyze here the worst case computational complexity for each one of the two-step configurations indicated in Table 1. Let us start by considering the simplest, most general non-adaptive batch statistical CS, where the dictionary (GMM) is learned and $M \ll N$ random measurements are used. The CS complexity is the same for all the methods considered here, since they all use M measurements per signal. Since CS consists of an $M \times N$ matrix by $N \times 1$ vector multiplication, for each signal, the complexity of CS is $O(\kappa SMN)$, where S is the total number of signals and κ is the number of MAP-EM iterations until convergence. The complexity required to estimate the original signal $\mathbf{x}$ (E-step) is dominated by matrix multiplication in the E-step (Equation (5)) for each Gaussian, which is $O(\kappa SGMN^2)$. Now, the complexity of the M-step is dominated by the update of the PCA basis (Equation (2)), which is $O(\kappa GN^3)$, since there are G Gaussians. Hence, the overall computational complexity of a random block GMM SCS with dictionary (and CS matrix) learning is $O(\kappa SGMN^2) + O(\kappa GN^3) + O(\kappa SMN)$. Since $S \gg N$, the dominant time complexity is $O(\kappa GSMN^2)$. Of course, dictionary (GMM) learning needs to be done only once (offline), hence, the complexity for an already learned dictionary and random sensing plus decoding becomes $O(GSMN^2)$.

The simplest configuration in Table 1 uses random sensing in the first step and the non-adaptive optimal (MSE) sensing for the estimated Gaussian in the second step. In this case, the computational complexity in the first and second steps is dominated again by matrix multiplication in (5), which is $O(\kappa SGMN^2)$. Since reconstruction must be done twice (one on each step), the overall increase in computational cost with respect to single step random sensing is a factor of two, which is marginal, and this extra cost is offset by the improvement in the reconstructions obtained using the two-step framework (see Section V). On the other hand, the deterministic RIP-AB sensing matrix needs to be computed only once, at every iteration of the MAP-EM algorithm, hence, its computational cost is only $O(\kappa GN^3)$. Even more, the Wienner filter in Equation (5) is the same for all for all signals and has also a computational cost of $O(\kappa GN^3)$, hence, the overall computational of RIP-AB is $O(\kappa G(N^3 + SMN))$, accounting for the matrix (Wiener filter) by vector (CS signal) multiplications in Equation (5). Note that we cannot precompute the Wiener filter when using a random sensing matrix, since it changes for every new signal (as it is recommended for improved sensing results, e.g., [19]). The RIP-AB matrix in the first step and the G possible optimal (MSE) non-adaptive matrices in the second step, can all be computed offline, before the sensing begins.

Now, the computational complexity of IDA (assuming an steepest ascend approach, see Equation (30))

is given by $O(\chi K^3)$, corresponding to χ steepest ascend iterations, where the computational complexity of each iteration is given by the cost of inverting $K \times K$ matrices, plus matrix multiplication of the same size. Since $K \leq M$, the worst case time complexity of IDA is $O(\chi M^3)$. Given that the sensing matrix needs to be computed only once every MAP-EM iteration,[1], and the Wiener filters are the same for all signals, the computational complexity (including E-M steps) of IDA is given by $O(\kappa(\chi M^3 + GSMN))$, where χ is in the order of $10^3 - 10^4$.

Finally, the computational complexity of AIDA-SHT is given by $O(\kappa S \chi K^4)$, where the computational cost is dominated by K/b IDA-like steepest ascent iterations, for each signal (see Algorithm 1). Since $K \leq M$, the worst case scenario of AIDA-SHT is given by $O(\kappa S \chi M^4)$, for $b = 1$. In the second step, the optimal (MI) adaptive sensing matrix can be pre-computed for every possible length K of the AIDA-SHT sensing matrix obtained in the first step and for every Gaussian. The cost in the second step would be $O(GN^3)$, but needs to be done only once (assuming a dictionary already learned). Hence, AIDA-SHT plus the optimal (MI) adaptive sensing matrix in the second step has a complexity $O(\kappa S(\chi M^4 + GMN^2))$, since the Wienner filters change for each signal. Since χ is usually of the order of $10^3 - 10^4$ and S can be of the order of 10^5 for non-overlapping patches and 10^7 for overlapping patches (depending of course on the image size), the time complexity of AIDA-SHT in the first step imposes a significant extra computational cost and should be done offline if possible, with a previously learned dictionary.

Even tough the proposed AIDA-SHT is the most expensive computationally, it has a great theoretical justification, potentially improving classification and reconstruction accuracies (see Section V). Further work is necessary to reduce the time complexity of AIDA-SHT. One possibility is to theoretically estimate the expected minimum number of adaptive samples K in the first step, for a given GMM and probability of classification error P_e. Another possibility is to learn by Monte Carlo simulations or direct experimentation what is the number of samples required in the first step, for a given class of signals.

[1]Despite the fact that IDA is usually initialized with a random sensing matrix, IDA (and AIDA) might be run several times offline, producing a good solution that is fixed (deterministic) for all signals.

Non-Adaptive Statistical Compressive Sensing

(a) (b)

(c) (d)

Fig. 1. Image reconstructed from learned dictionaries and non-overlapping patches of size 8×8 (CS to 12 samples). a) Original, b) Random (23.2 dbs), c) Unstructured/Structured (23.3 dbs), d) RIP-AB (26.6 dbs).

(a) (b)

(c) (d)

Fig. 2. Image reconstructed from learned dictionaries and non-overlapping patches of size 8×8 (CS to 12 samples). a) Original, b) Random (23.9 dbs), c) Unstructured/Structured (23.9 dbs), d) RIP-AB (26.4 dbs).

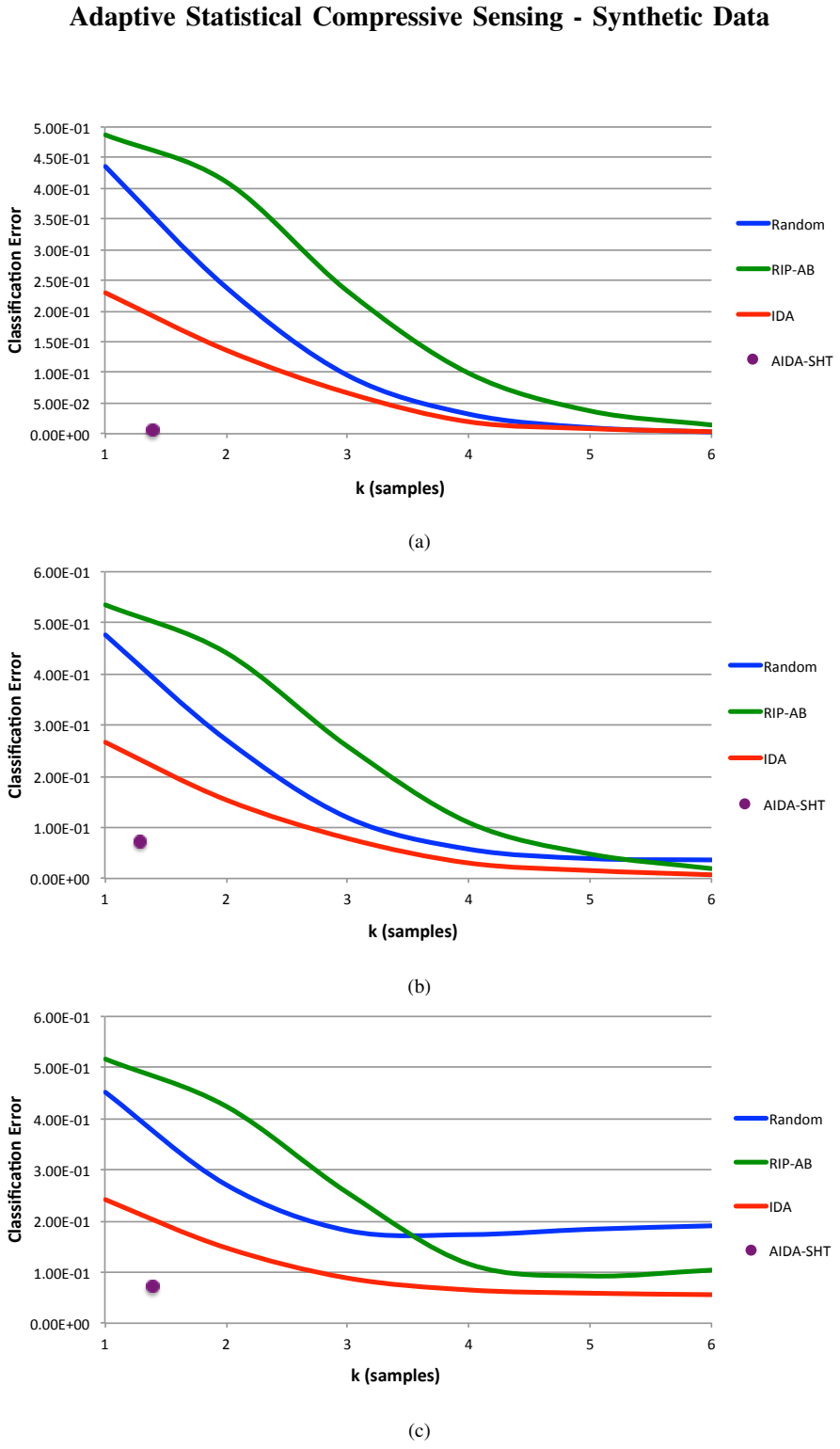
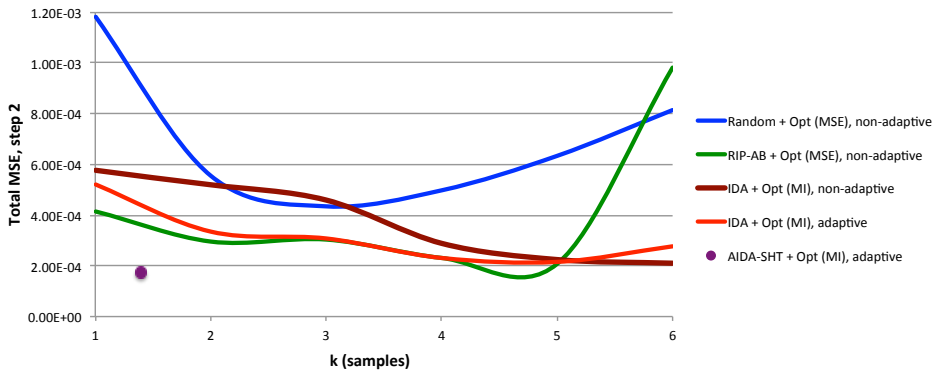
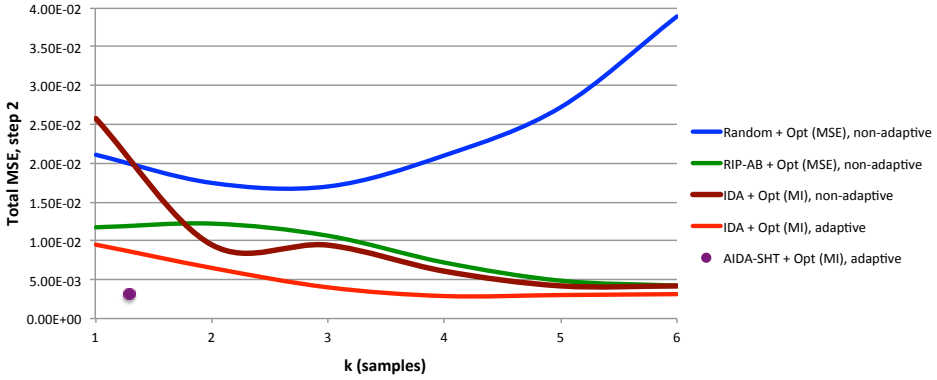


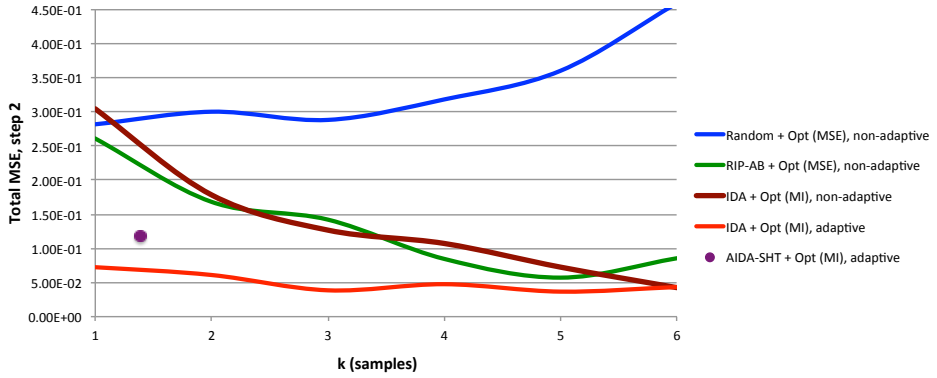
Fig. 3. Classification accuracy (step 1) synthetic signals of dimension 36 (CS to 6 samples) BD $\in [30\ 46)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

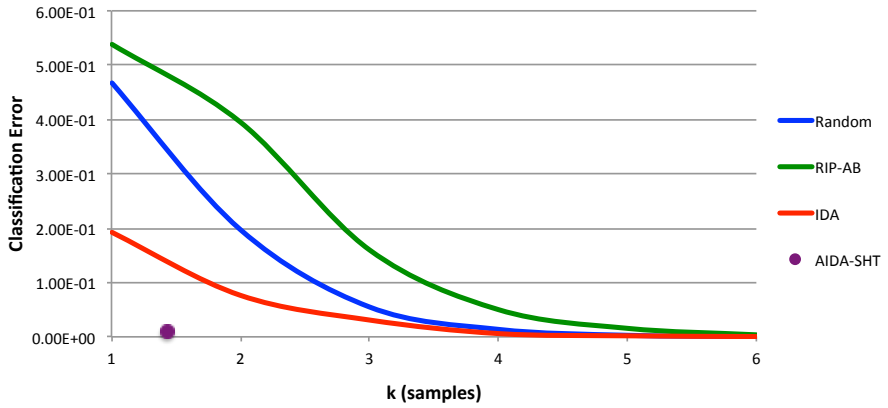


(b)

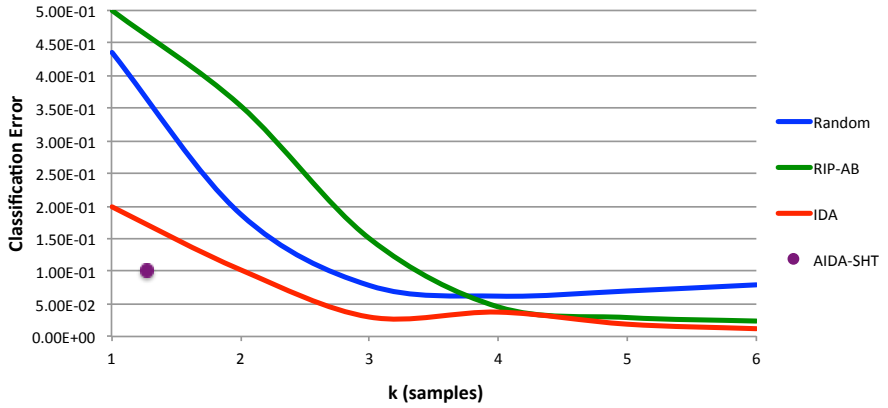


(c)

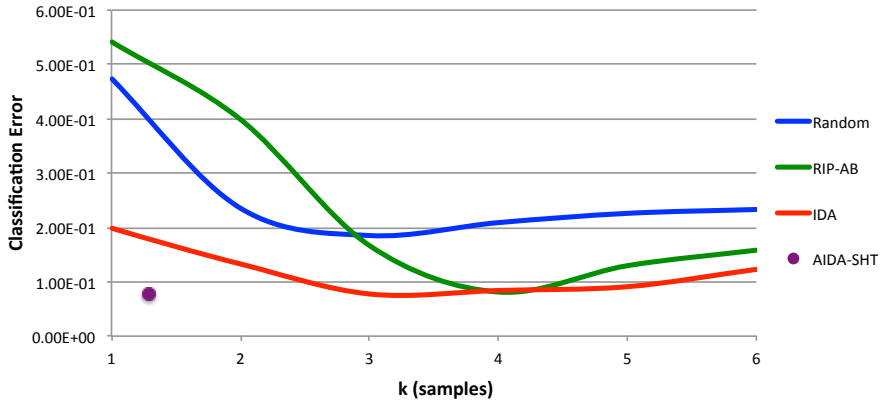
Fig. 4. MSE (step 2) reconstructed synthetic signals of dimension 36 (CS to 6 samples) BD $\in [30\ 46)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

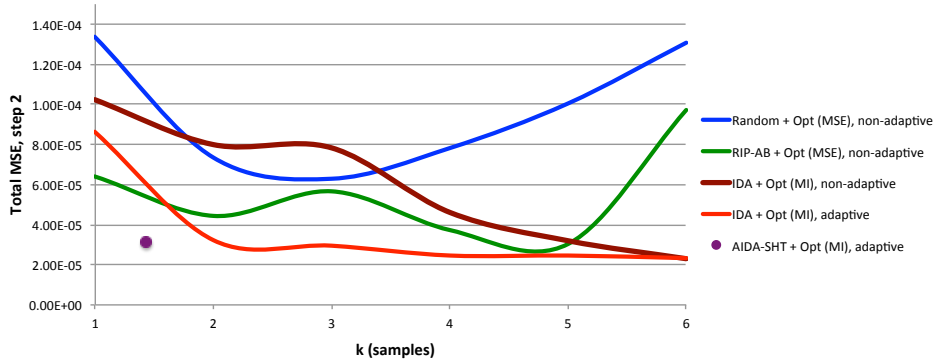


(b)

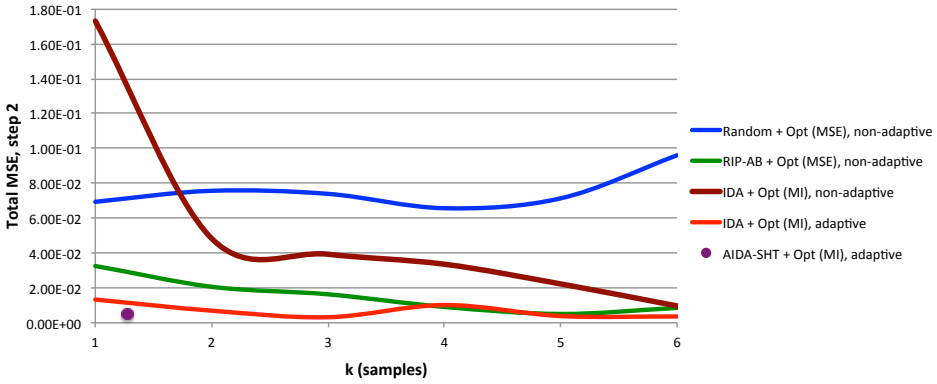


(c)

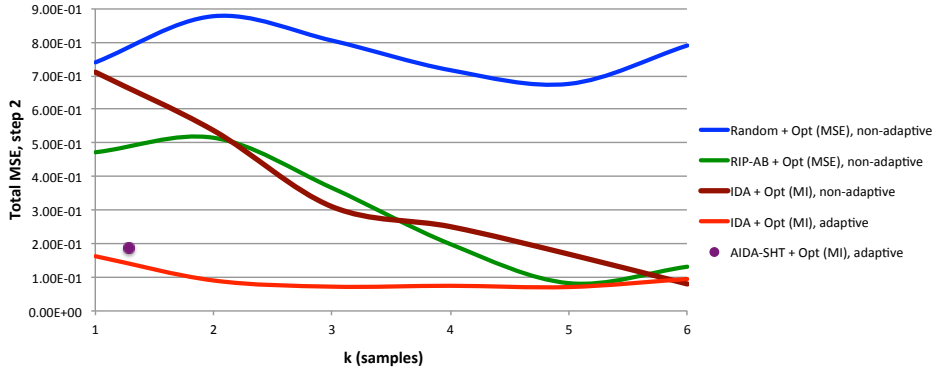
Fig. 5. Classification accuracy (step 1) synthetic signals of dimension 36 (CS to 6 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

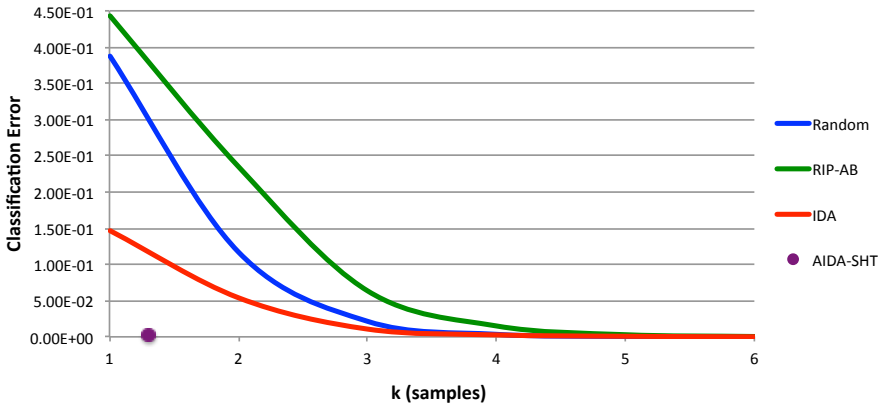


(b)

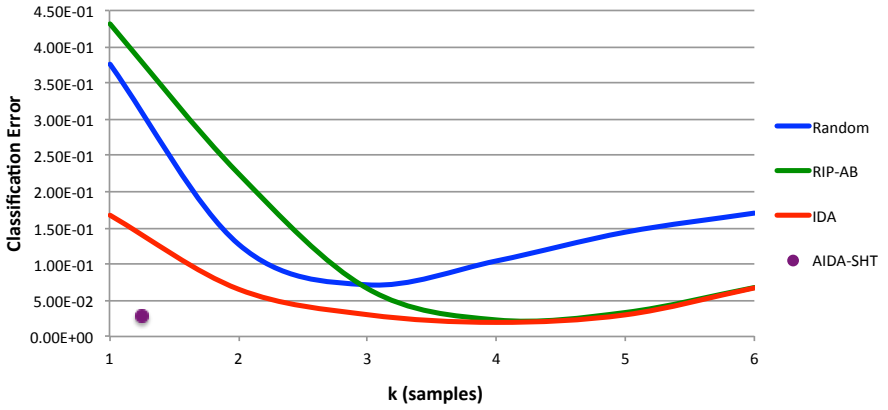


(c)

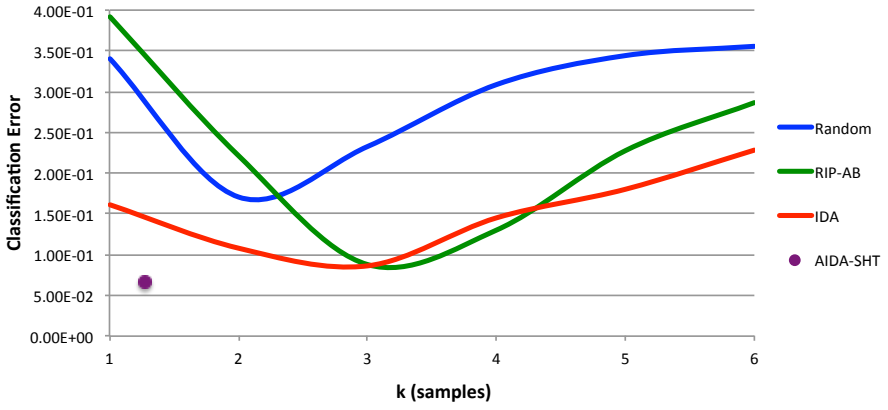
Fig. 6. MSE (step 2) reconstructed synthetic signals of dimension 36 (CS to 6 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

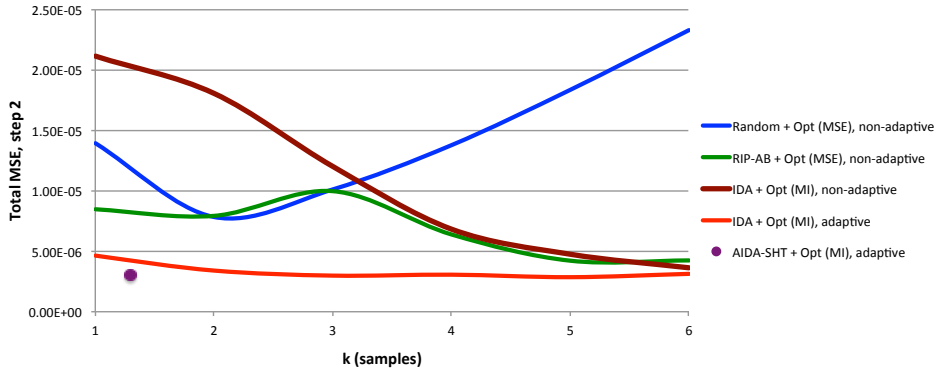


(b)

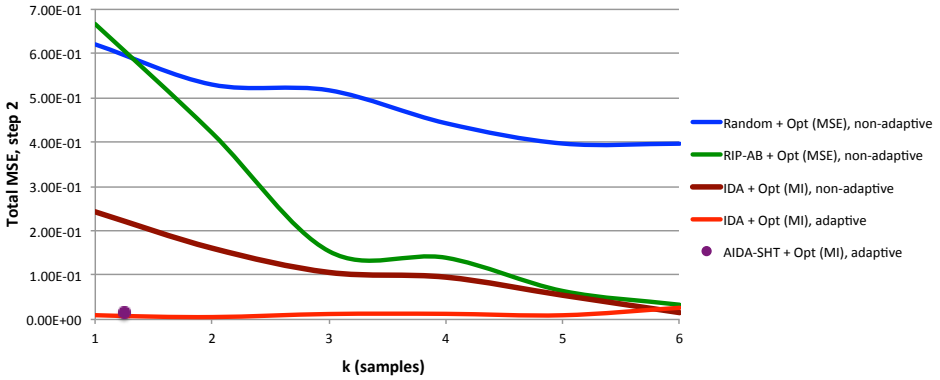


(c)

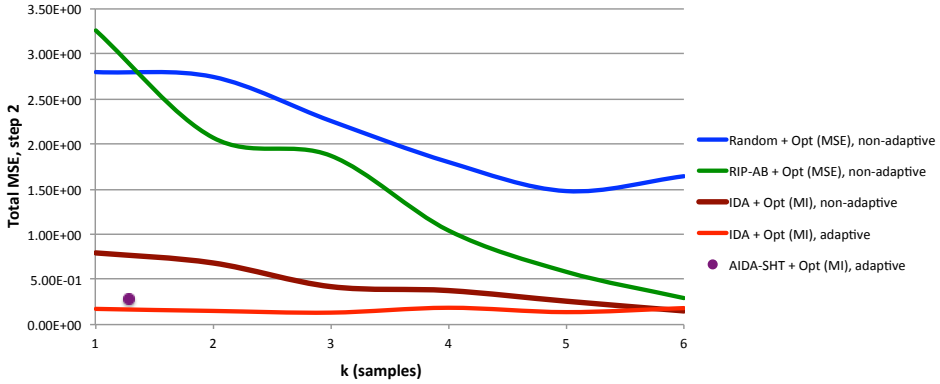
Fig. 7. Classification accuracy (step 1) synthetic signals of dimension 36 (CS to 6 samples) BD $\in$ [62 78). a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

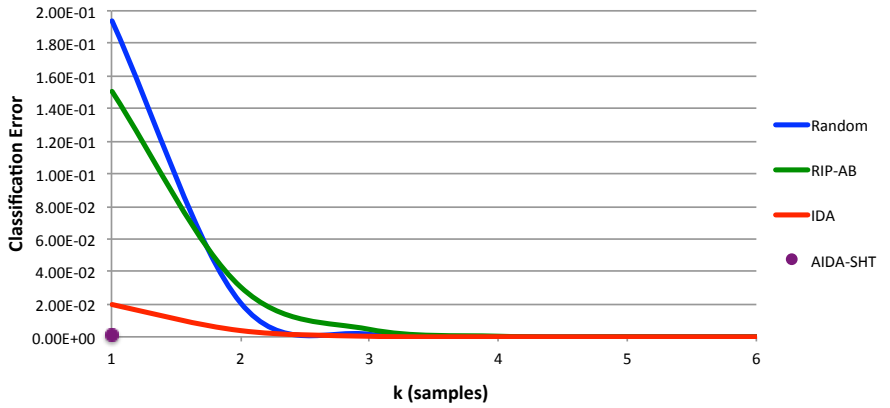


(b)

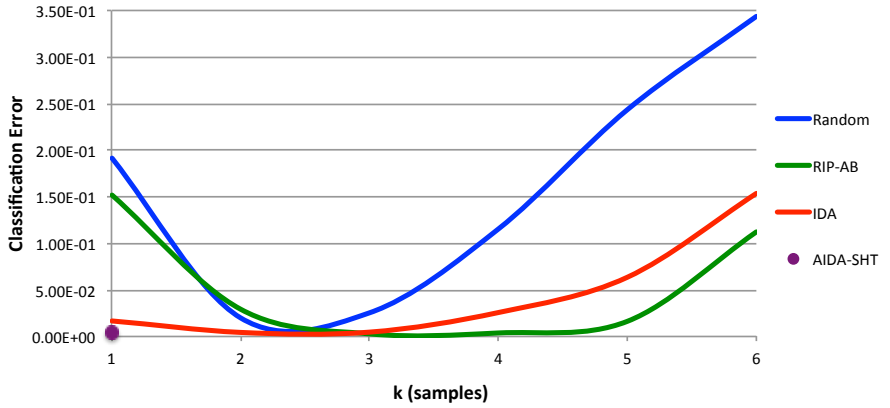


(c)

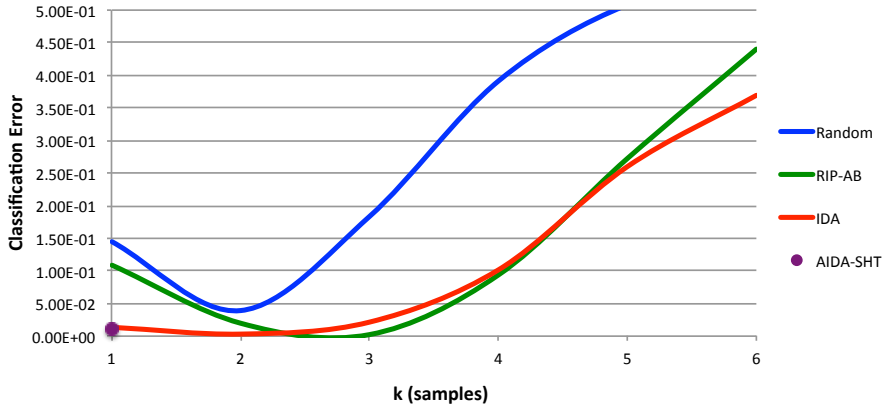
Fig. 8. MSE (step 2) reconstructed synthetic signals of dimension 36 (CS to 6 samples) BD $\in [62\ 78)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

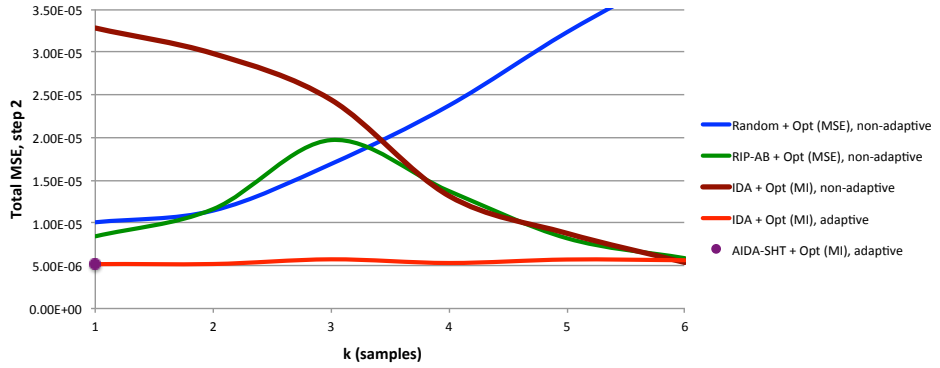


(b)

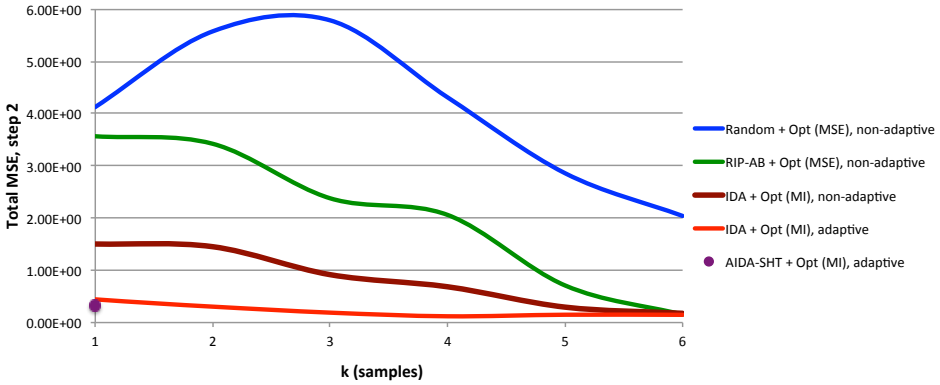


(c)

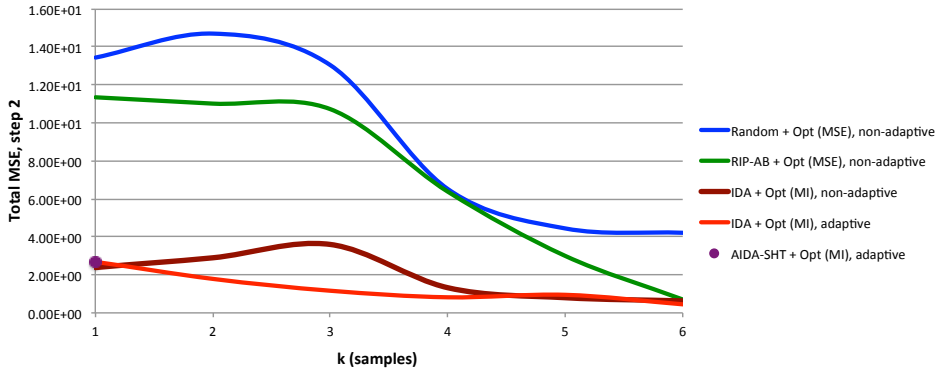
Fig. 9. Classification accuracy (step 1) synthetic signals of dimension 36 (CS to 6 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

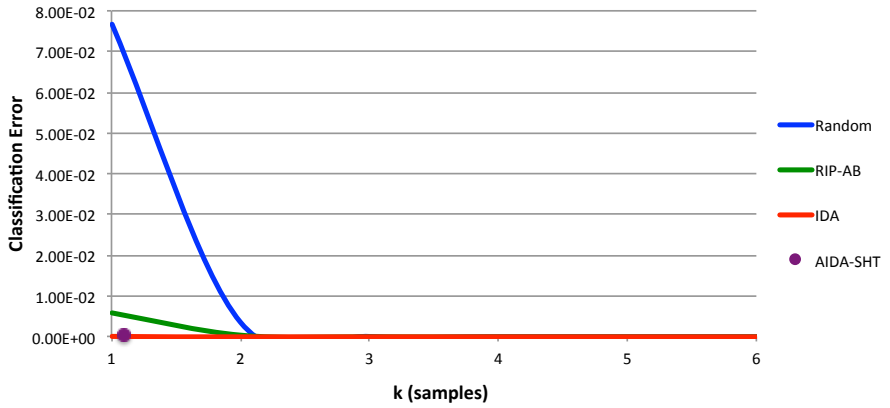


(b)

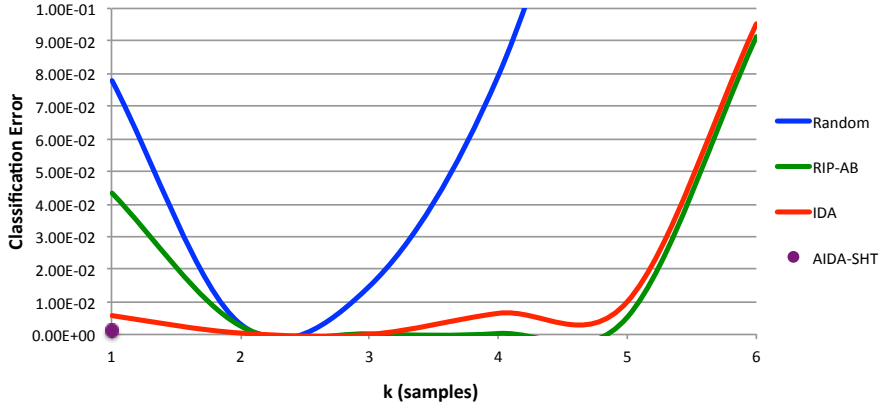


(c)

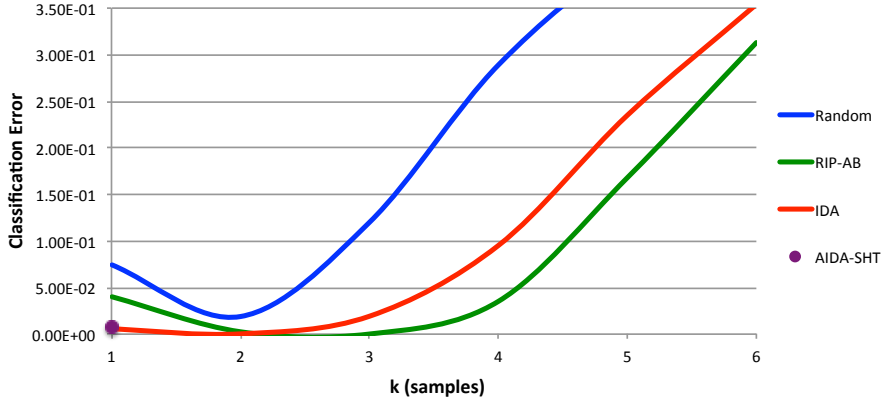
Fig. 10. MSE (step 2) reconstructed synthetic signals of dimension 36 (CS to 6 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

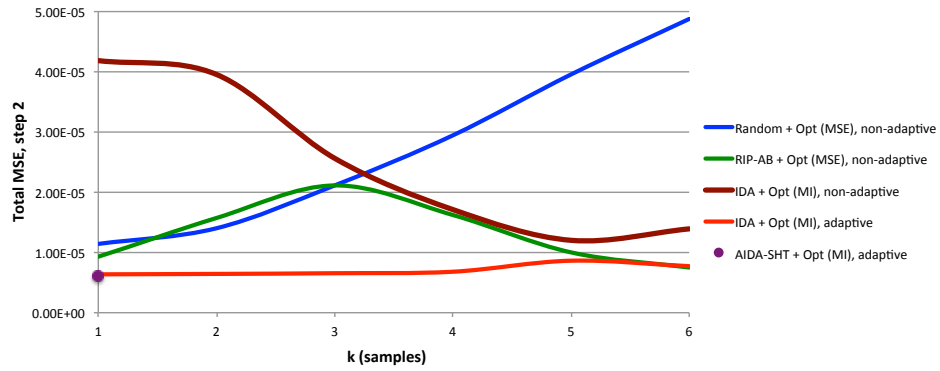


(b)

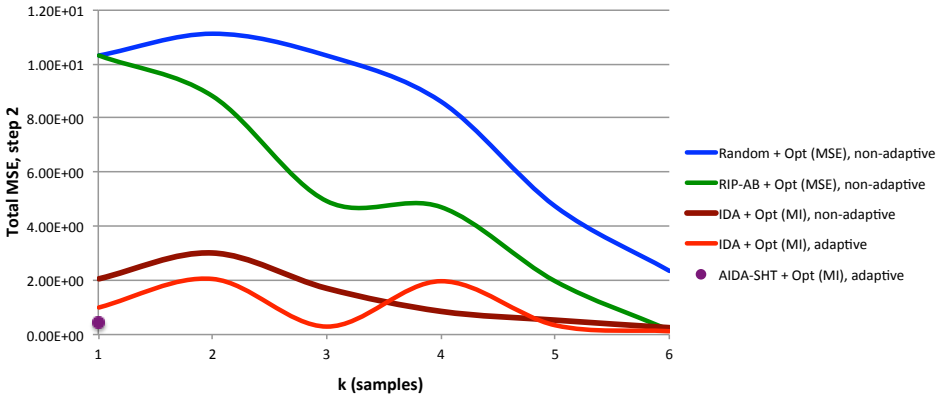


(c)

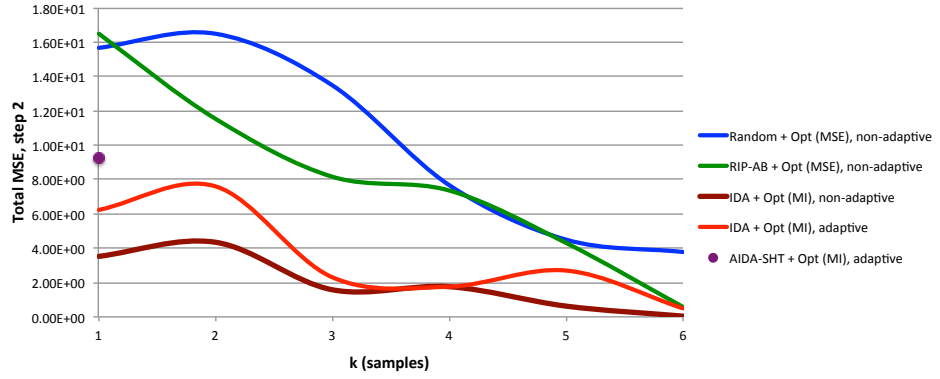
Fig. 11. Classification accuracy (step 1) synthetic signals of dimension 36 (CS to 6 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

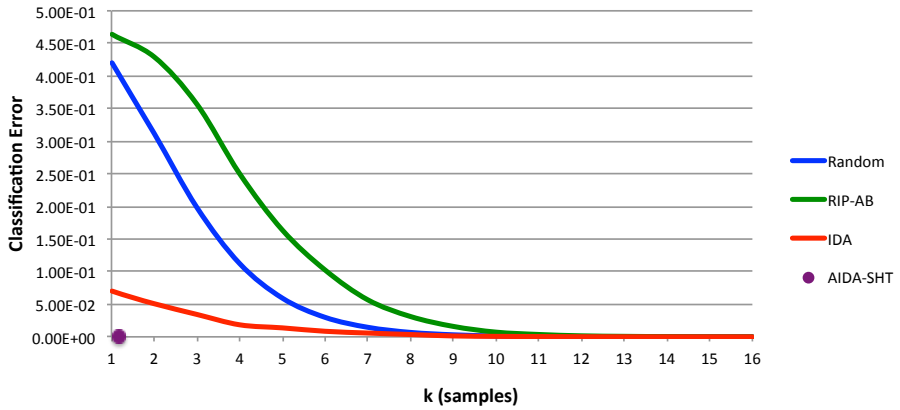


(b)

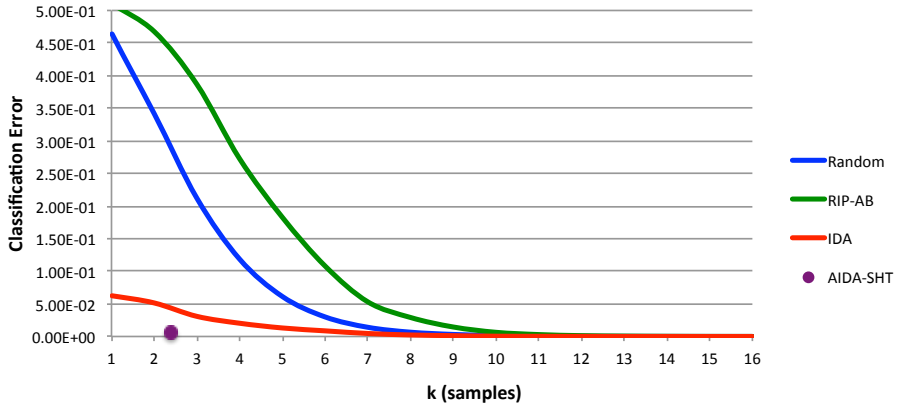


(c)

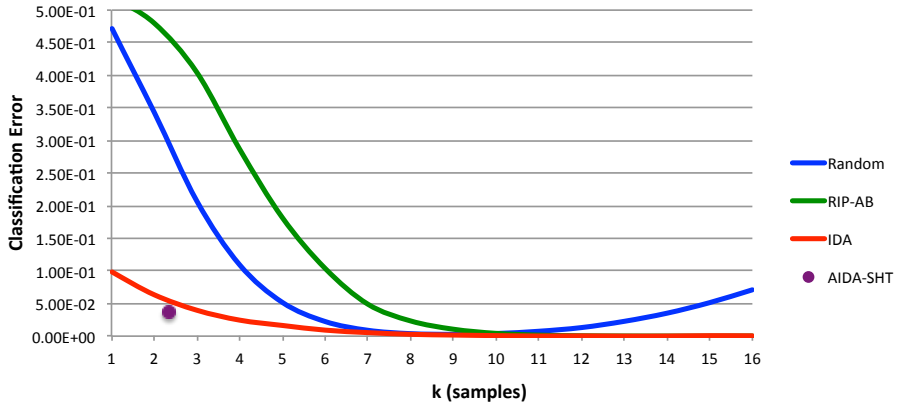
Fig. 12. MSE (step 2) reconstructed synthetic signals of dimension 36 (CS to 6 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

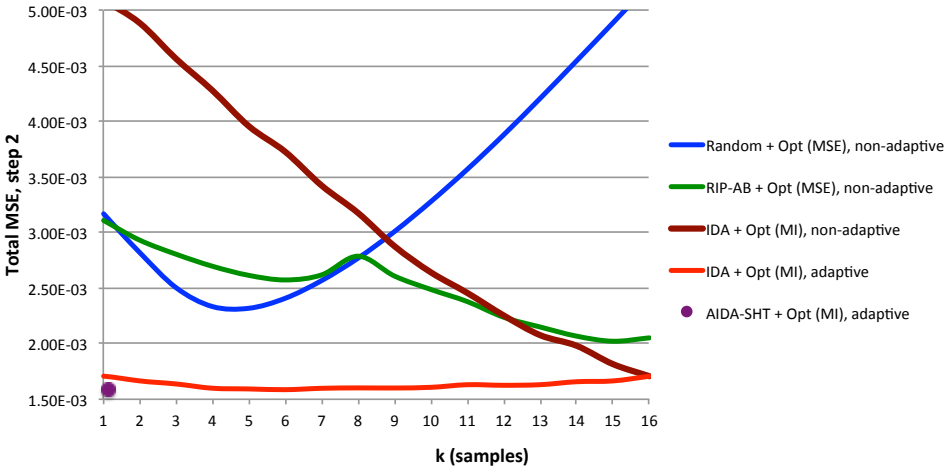


(b)

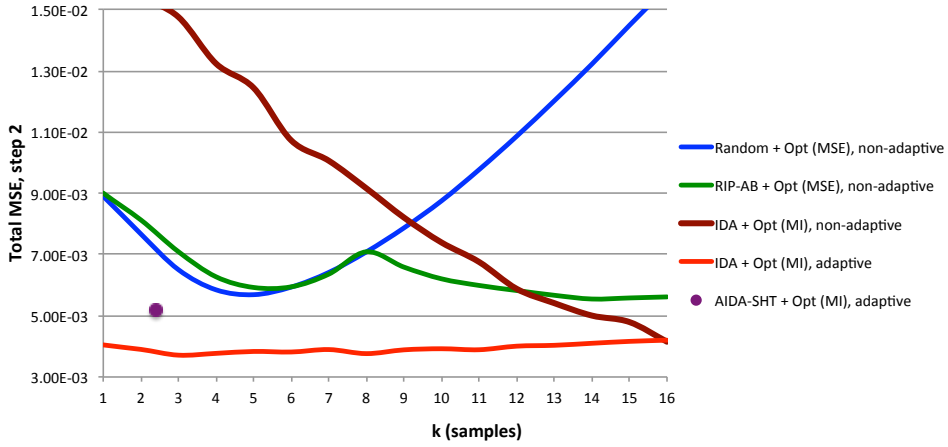


(c)

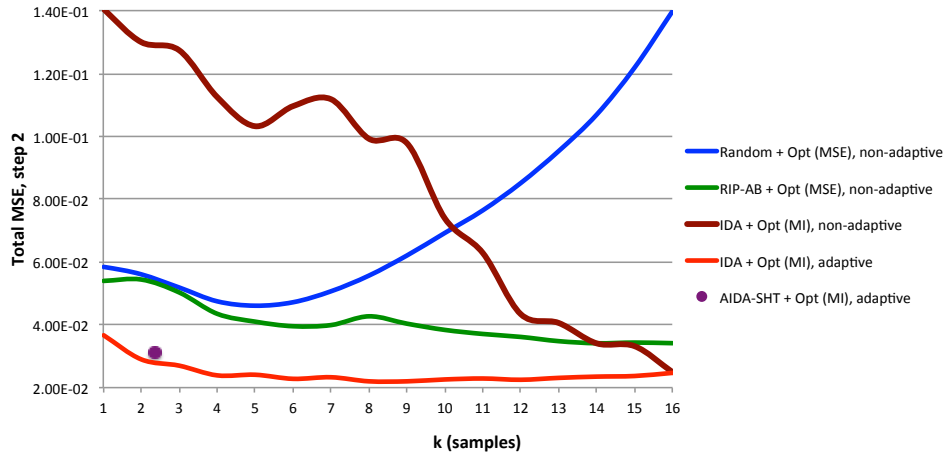
Fig. 13. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

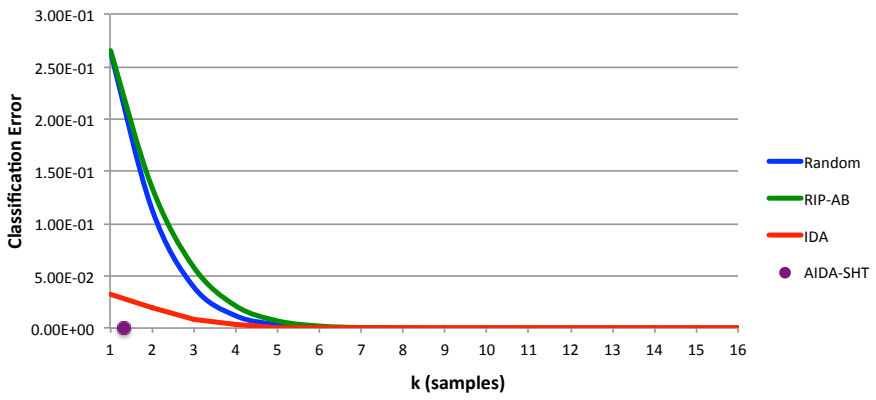


(b)

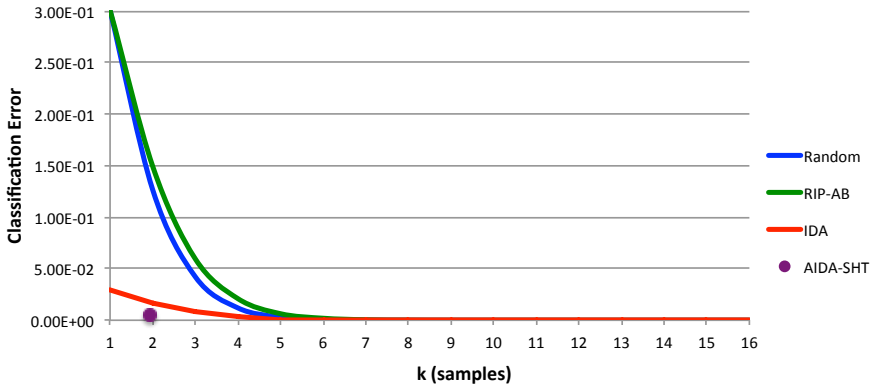


(c)

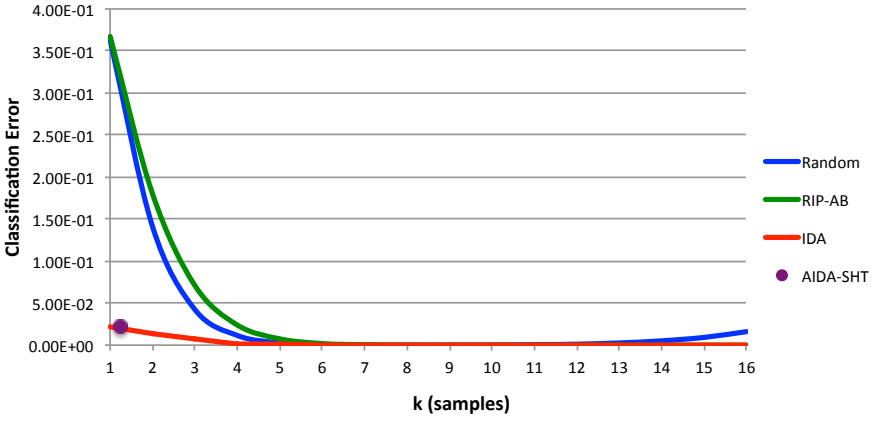
Fig. 14. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

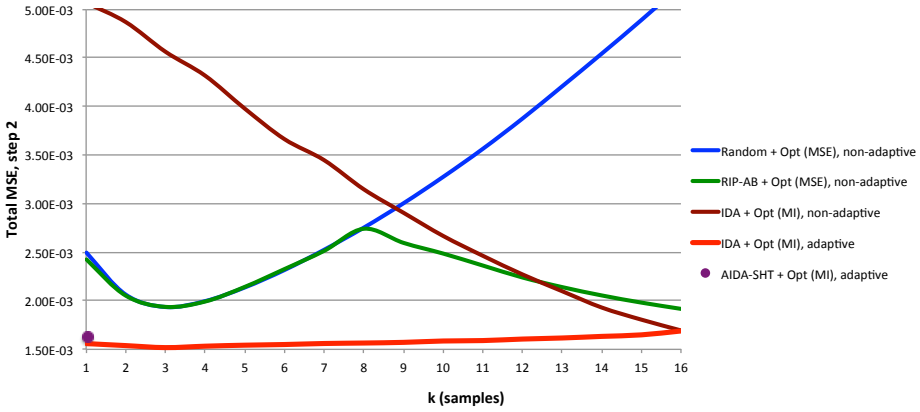


(b)

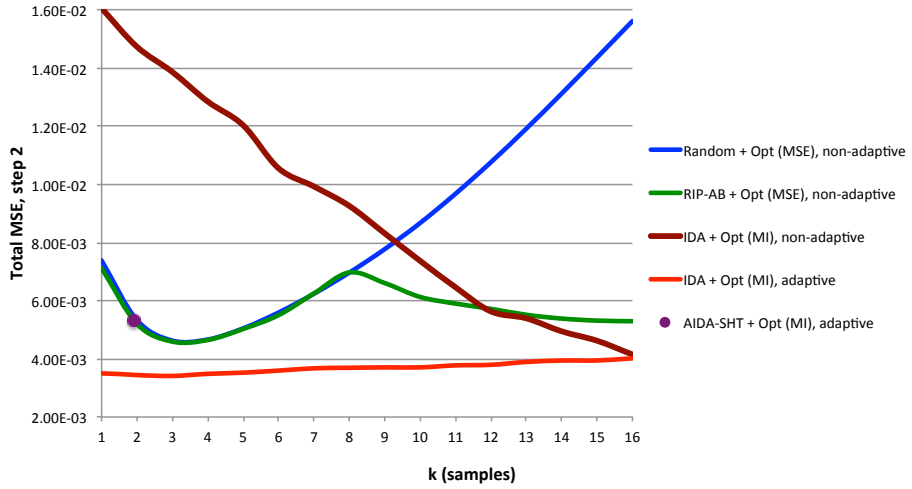


(c)

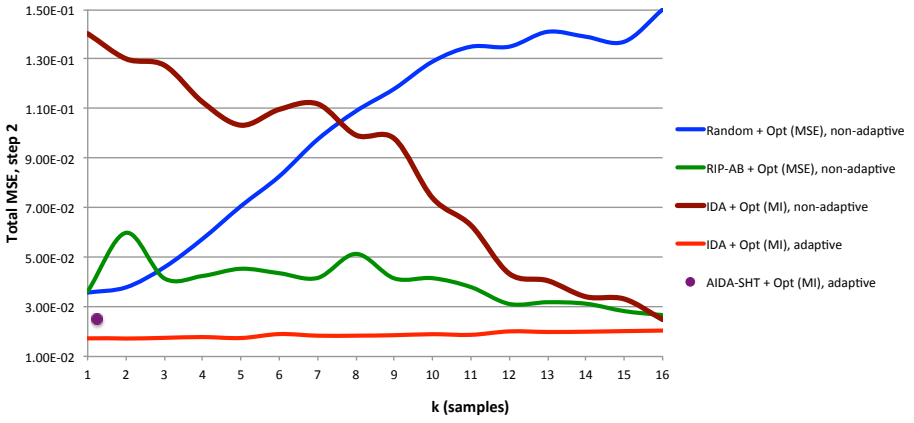
Fig. 15. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

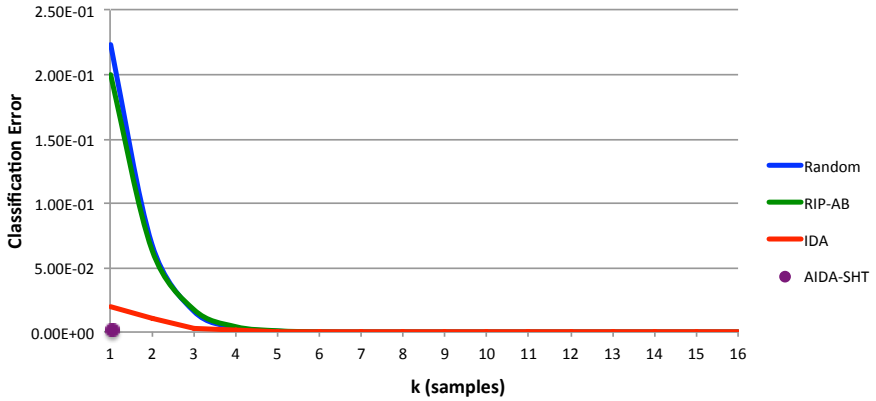


(b)

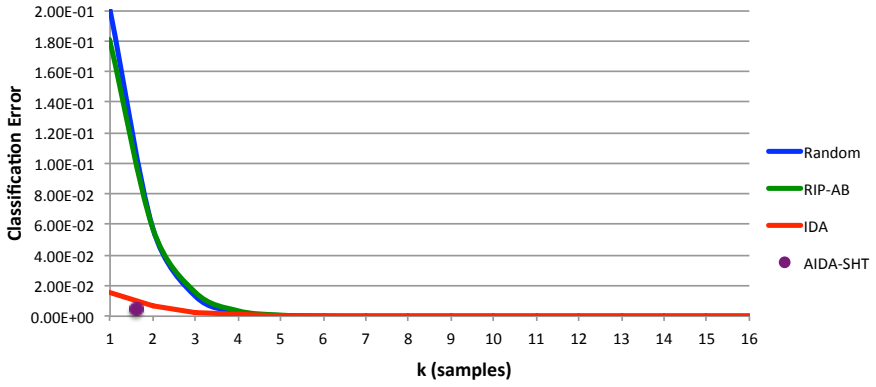


(c)

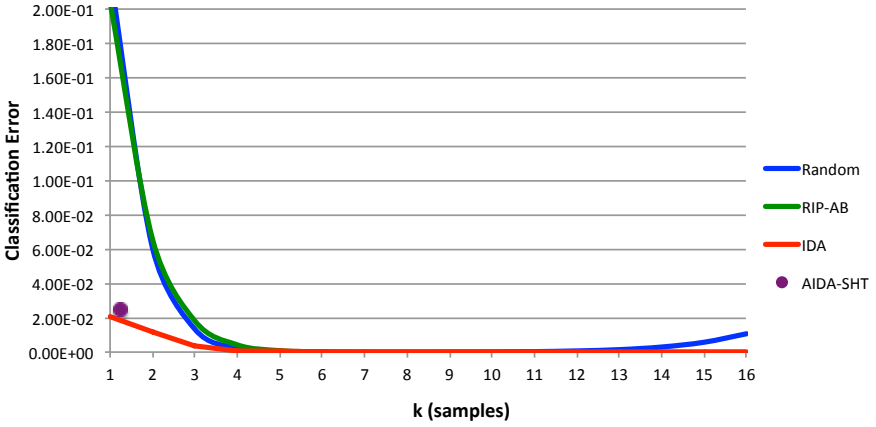
Fig. 16. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

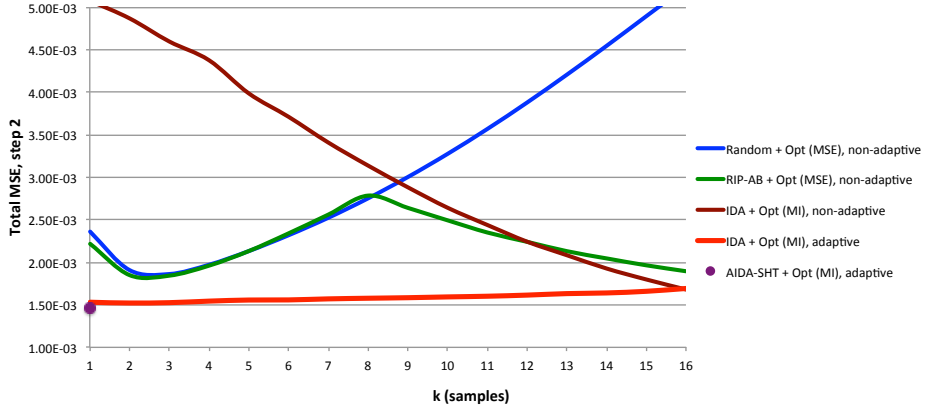


(b)

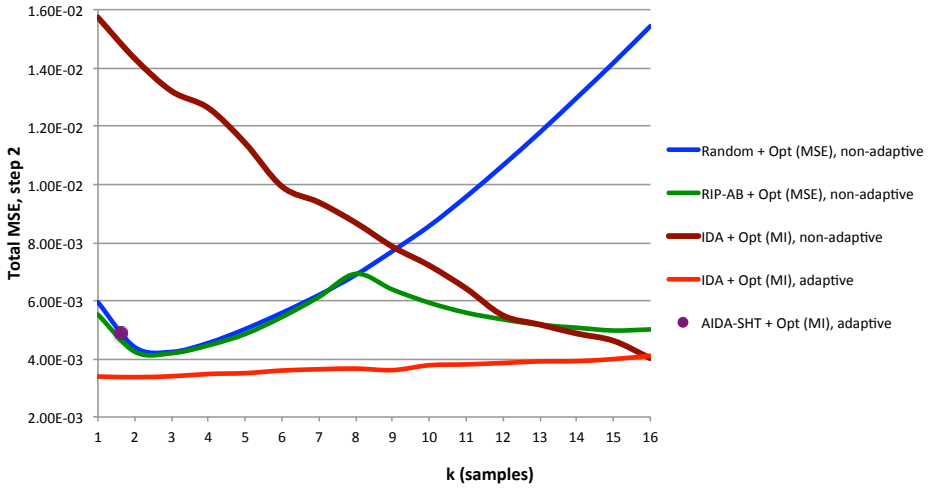


(c)

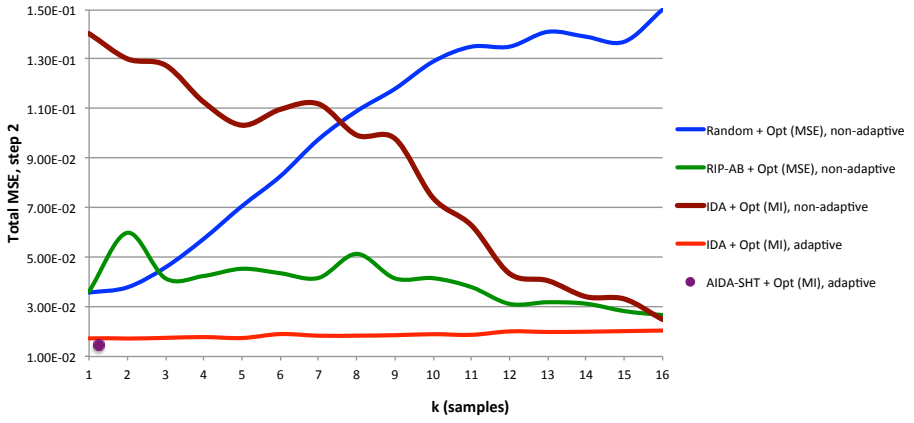
Fig. 17. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

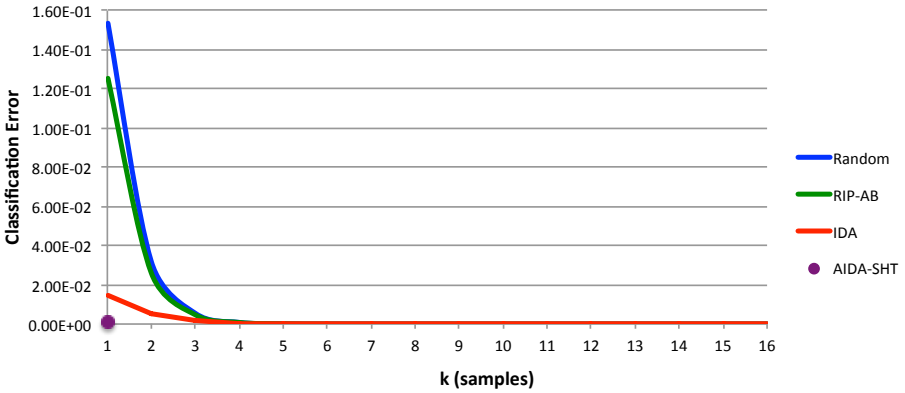


(b)

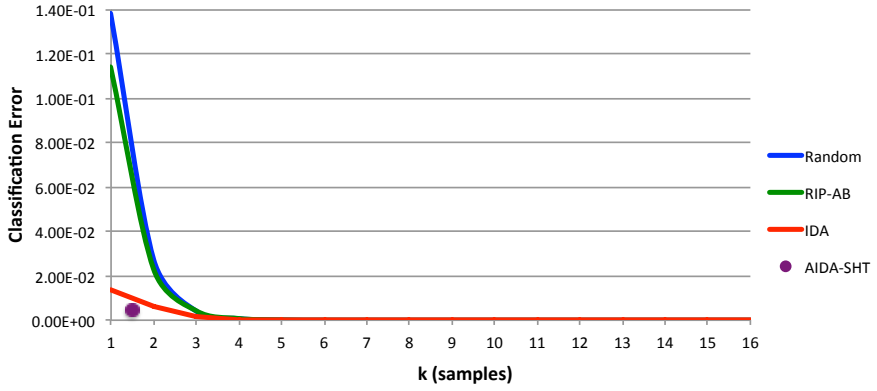


(c)

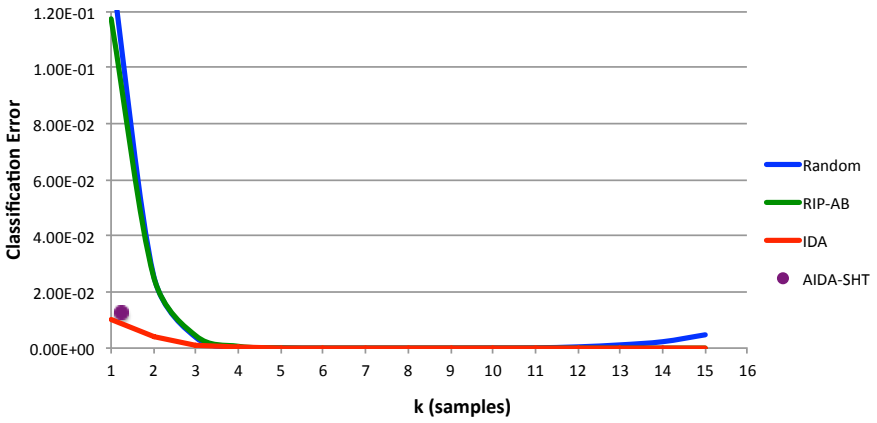
Fig. 18. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

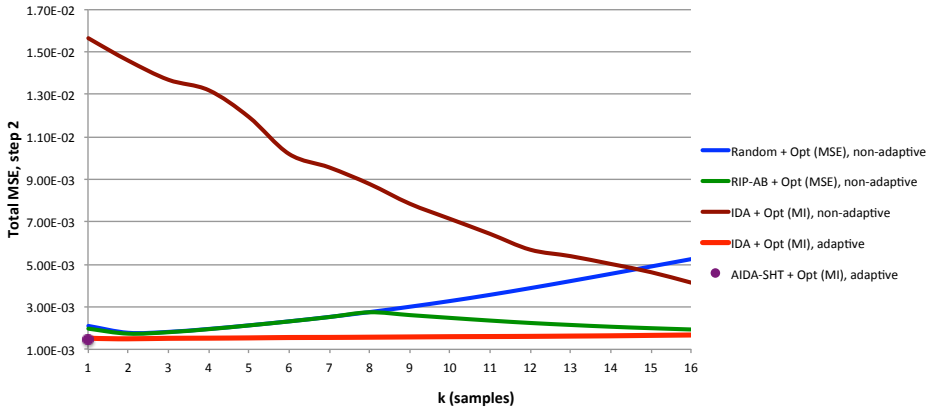


(b)

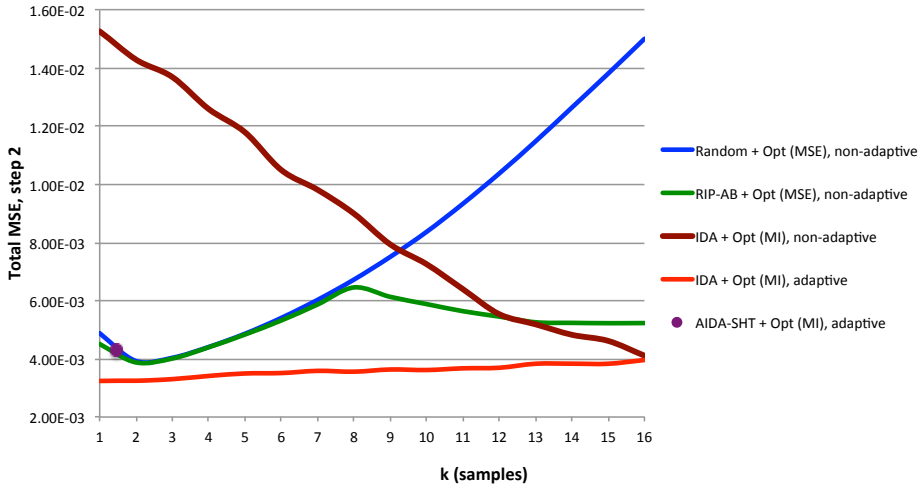


(c)

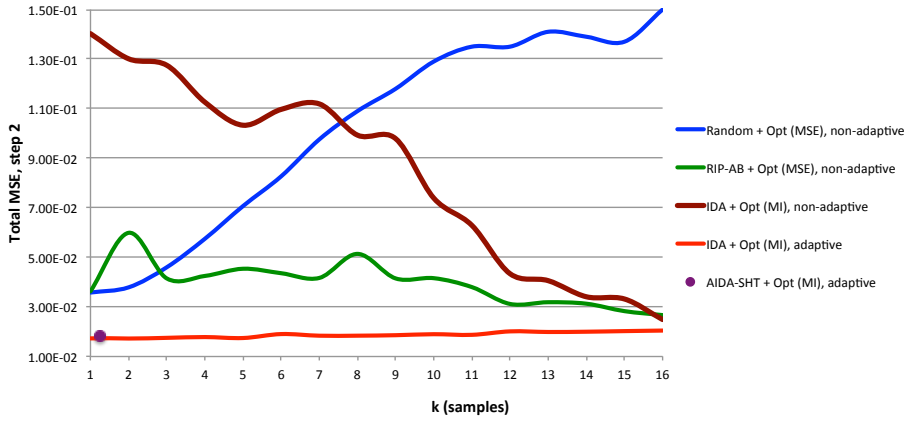
Fig. 19. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [110\ 126)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

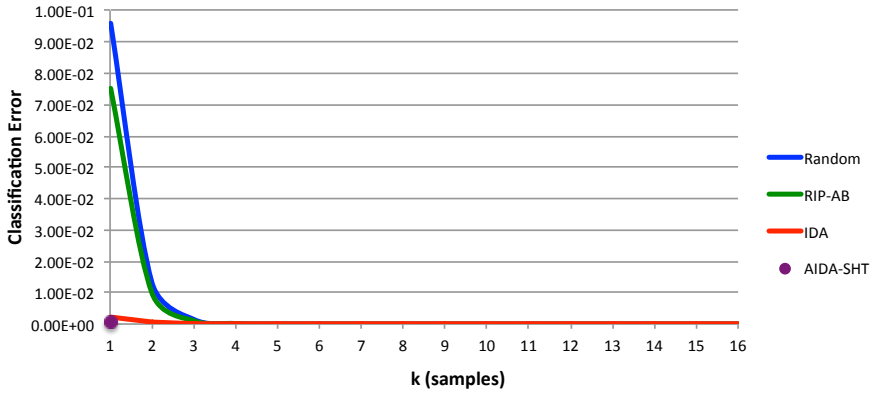


(b)

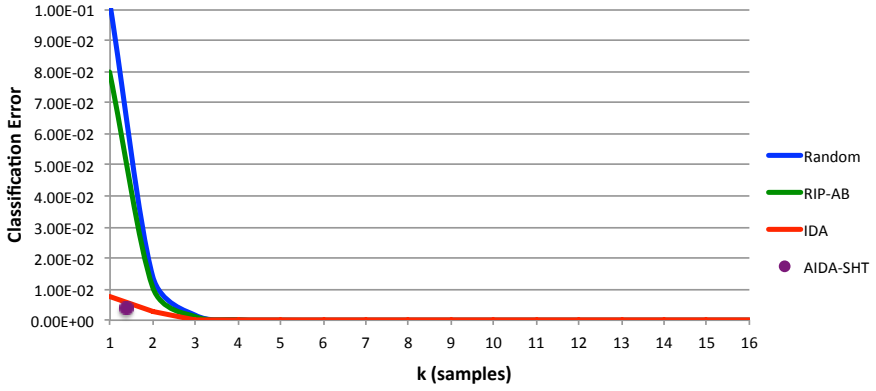


(c)

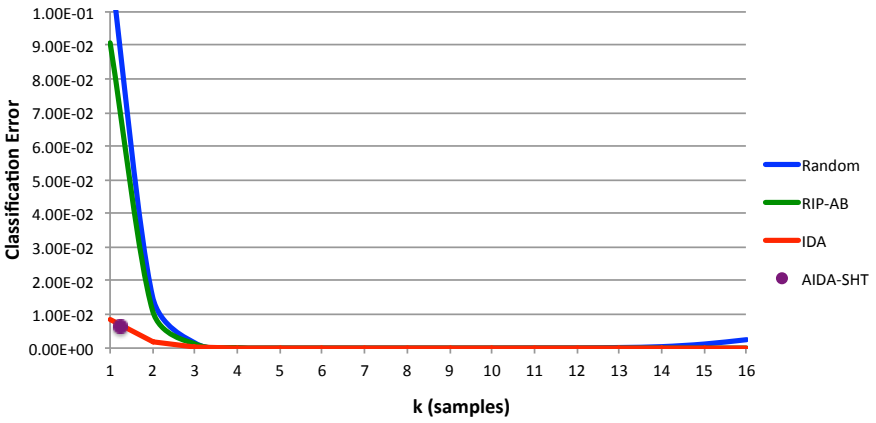
Fig. 20. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [110\ 126)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

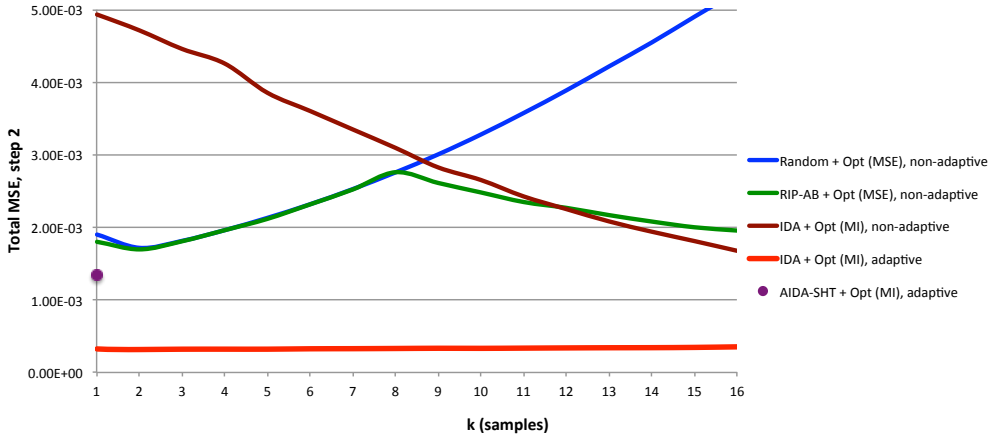


(b)

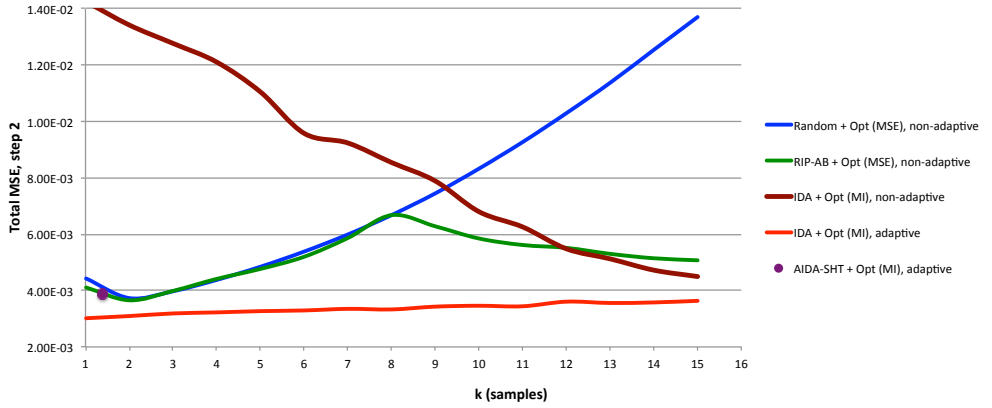


(c)

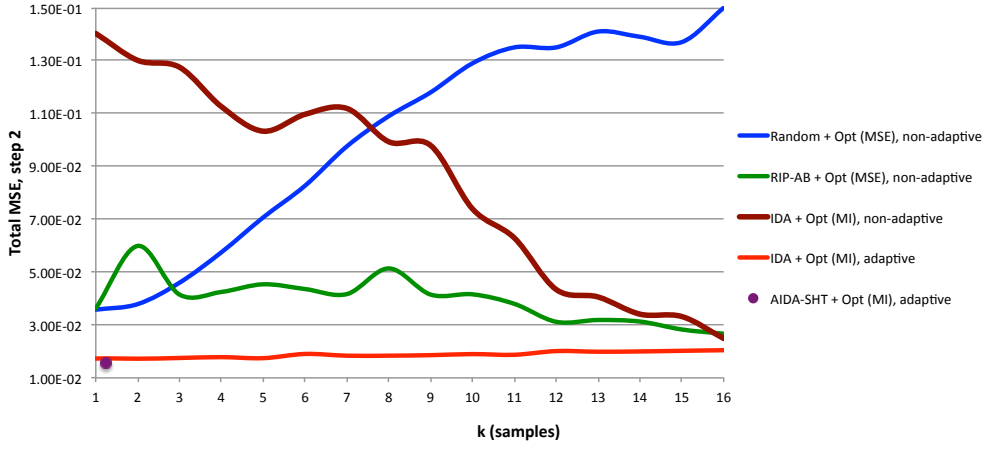
Fig. 21. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [126\ 142)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

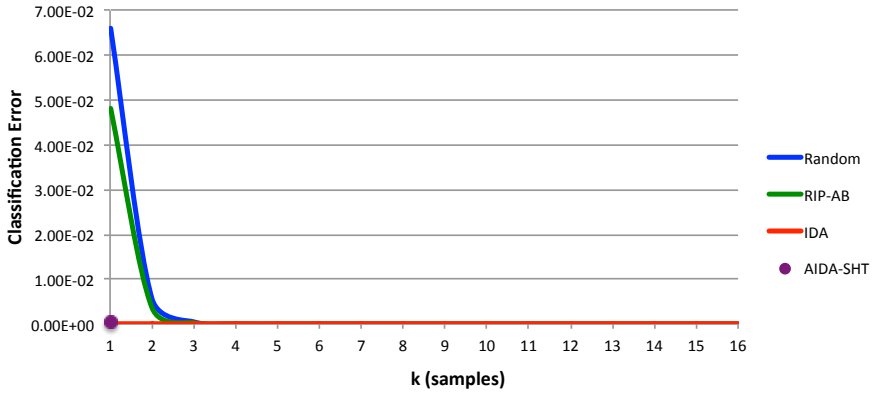


(b)

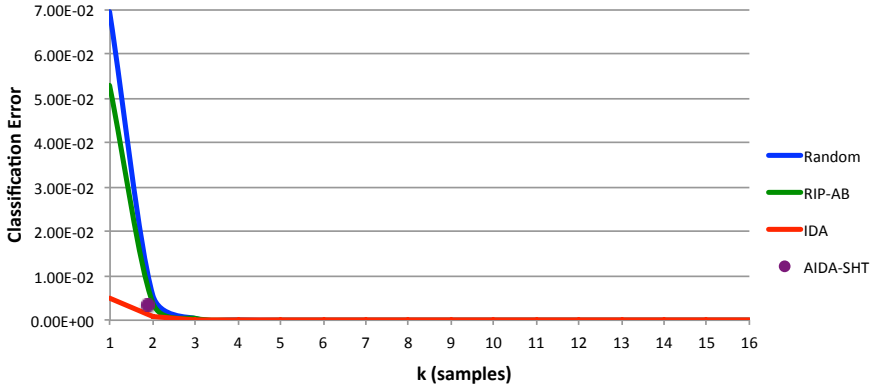


(c)

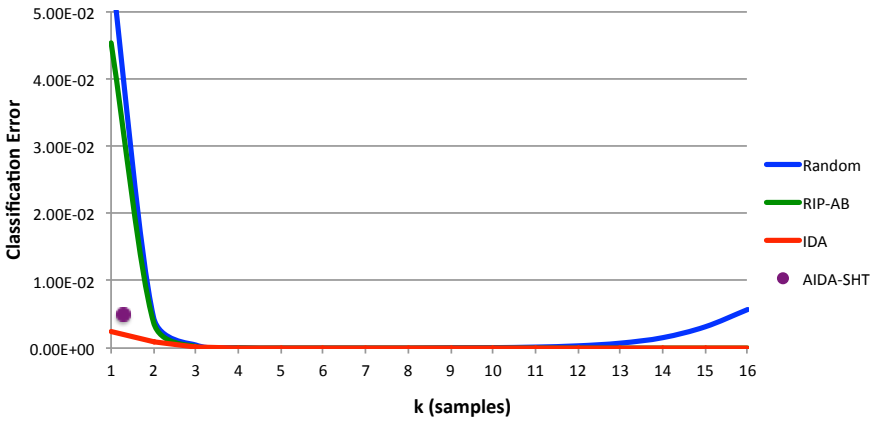
Fig. 22. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [126\ 142)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

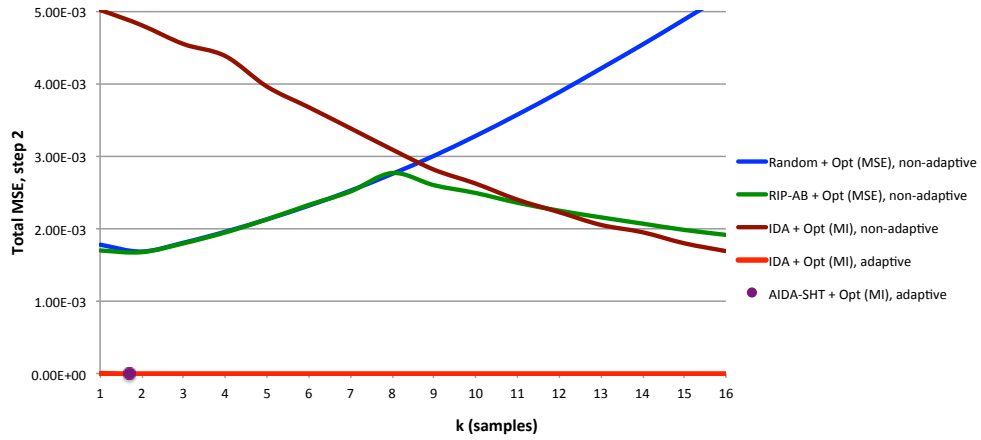


(b)

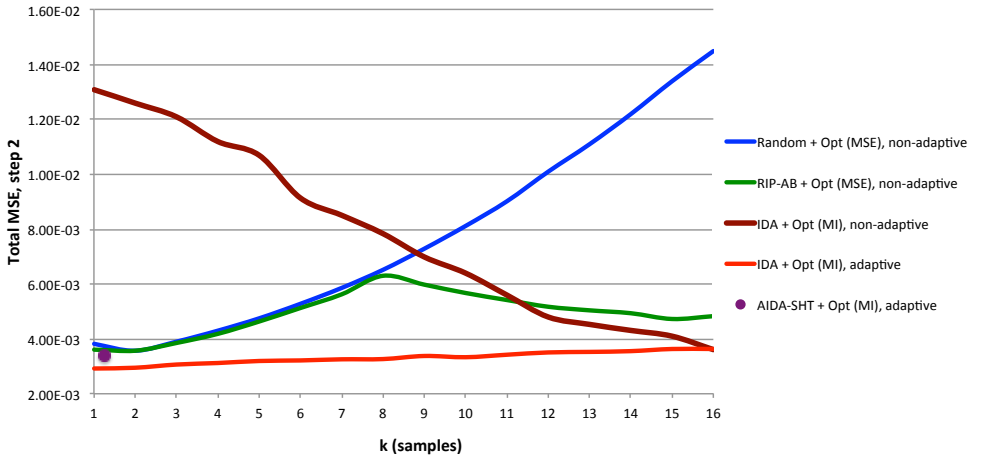


(c)

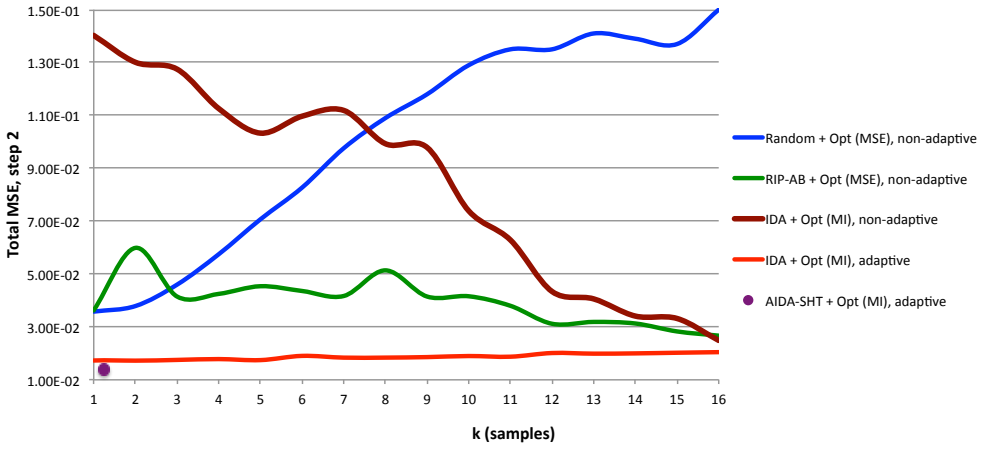
Fig. 23. Classification accuracy (step 1) synthetic signals of dimension 64 (CS to 16 samples) BD $\in [142 + \infty)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

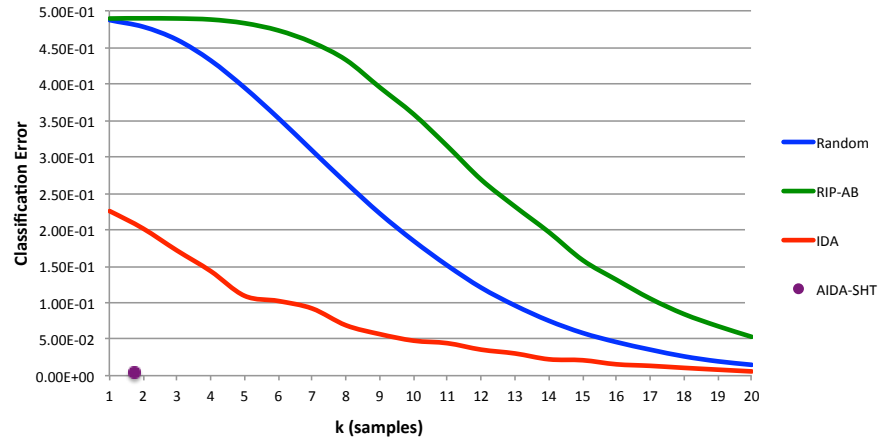


(b)

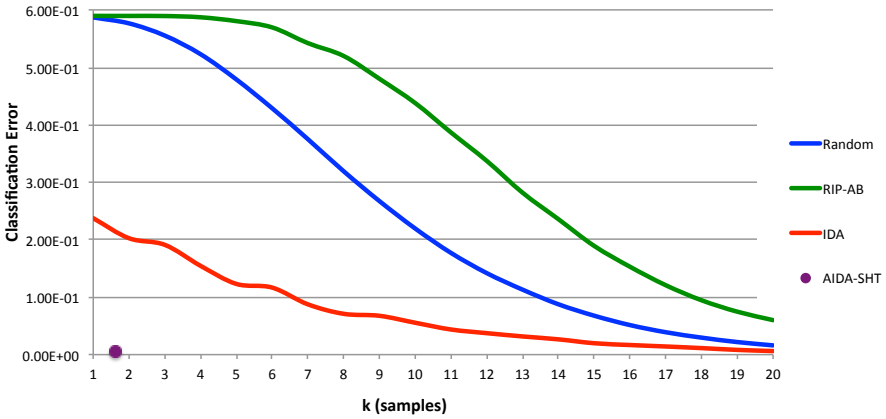


(c)

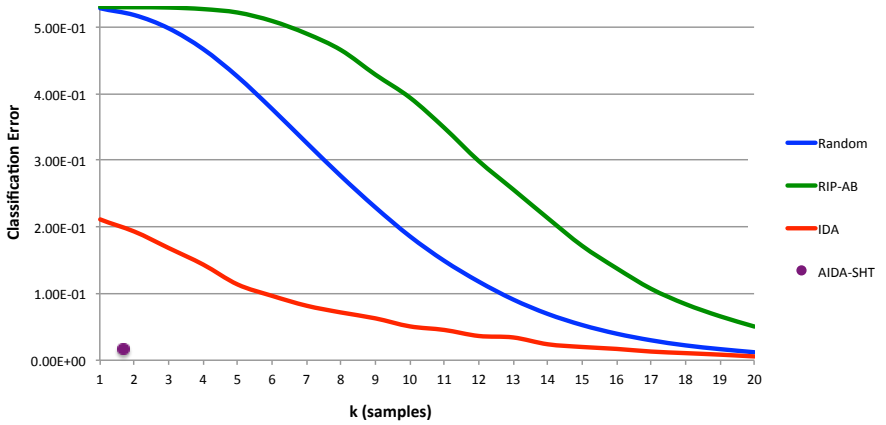
Fig. 24. MSE (step 2) reconstructed synthetic signals of dimension 64 (CS to 16 samples) BD $\in [142 \ +\infty)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

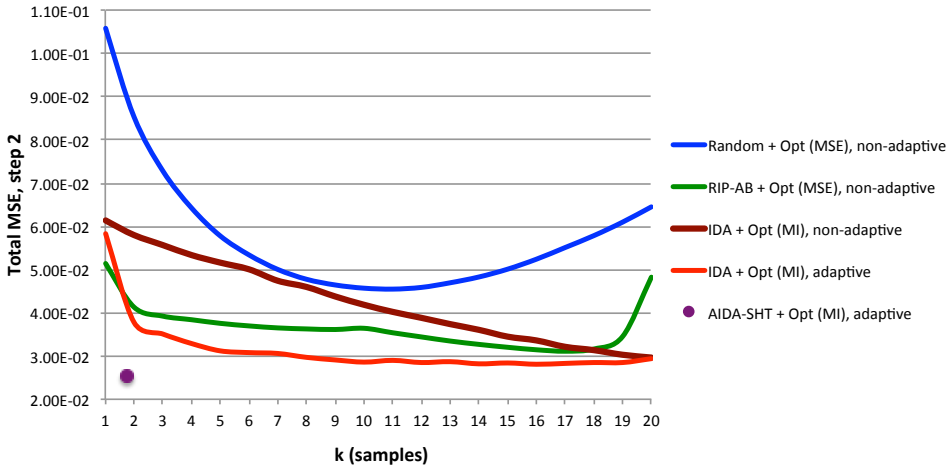


(b)

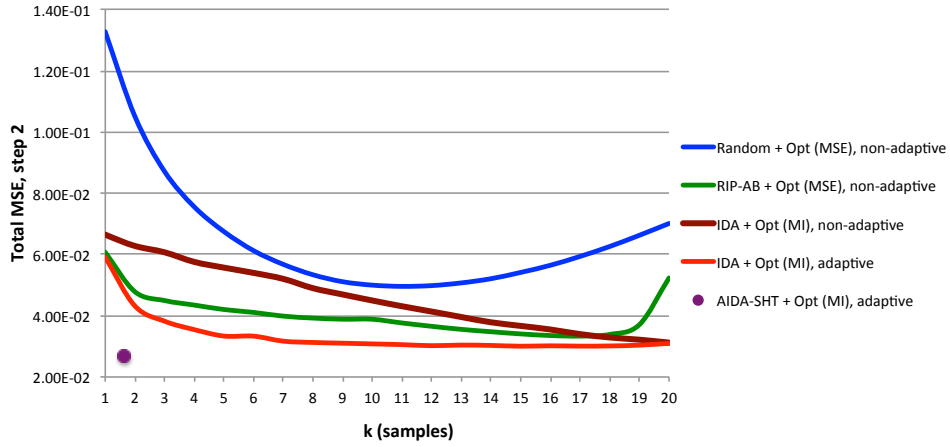


(c)

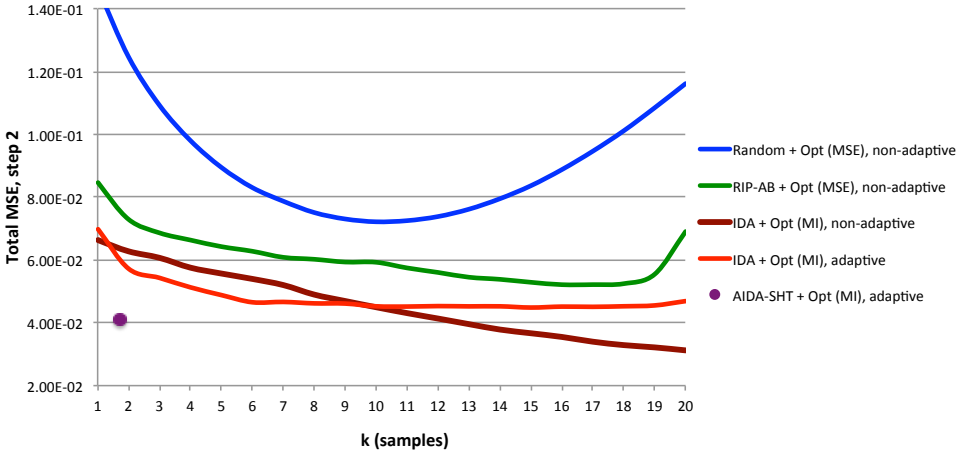
Fig. 25. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [30\ 46)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

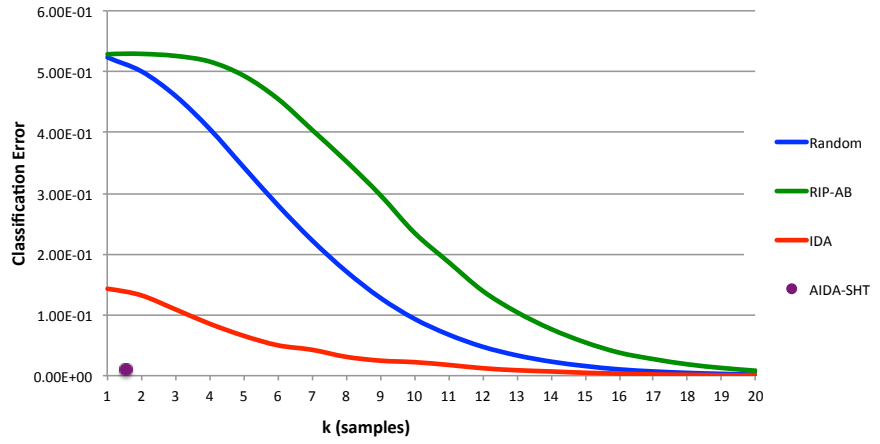


(b)

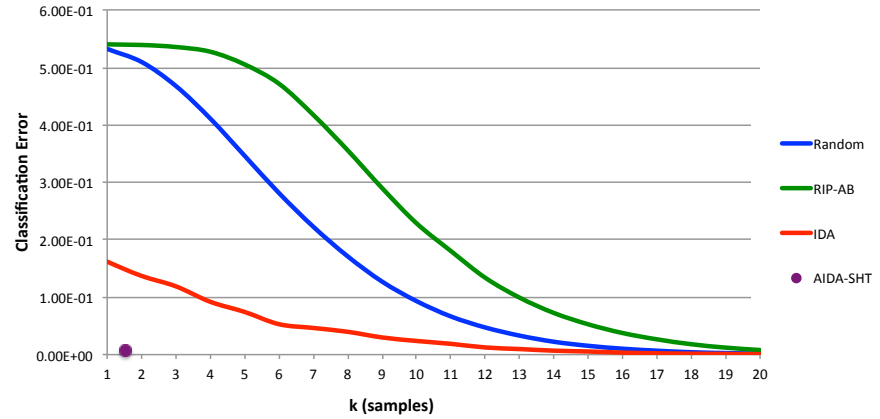


(c)

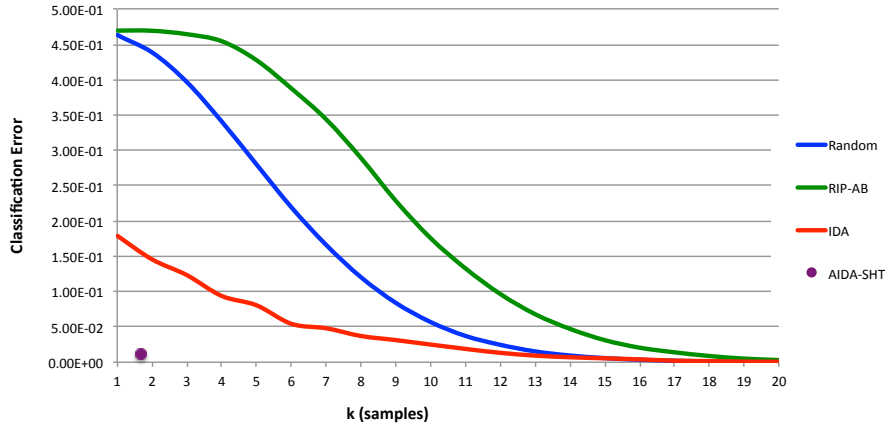
Fig. 26. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [30\ 46)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

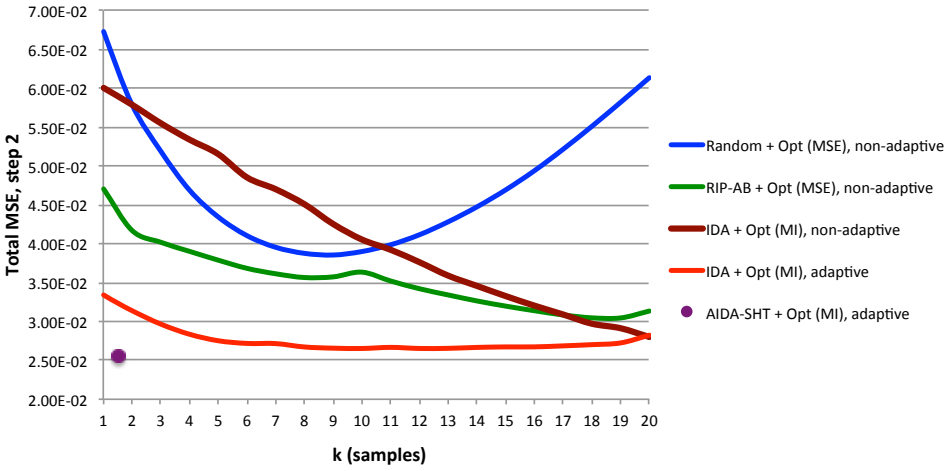


(b)

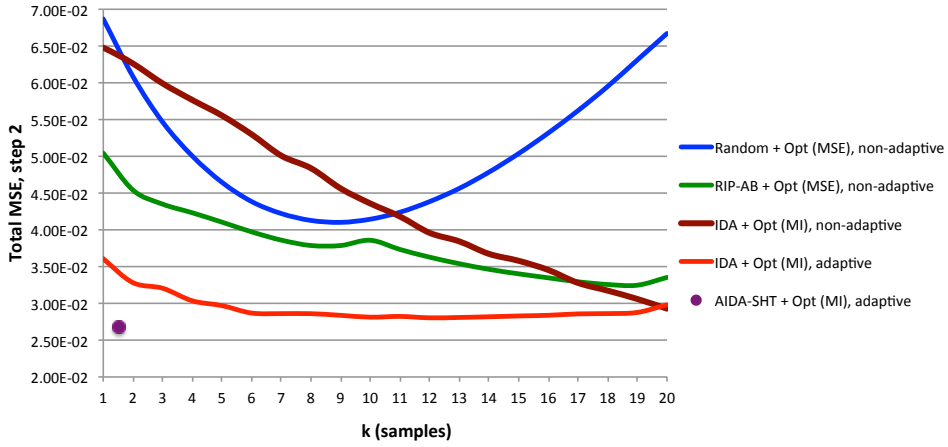


(c)

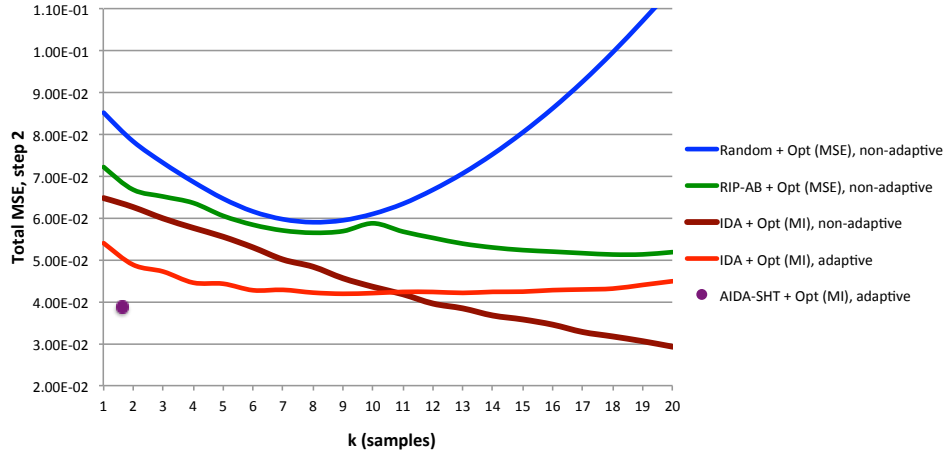
Fig. 27. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

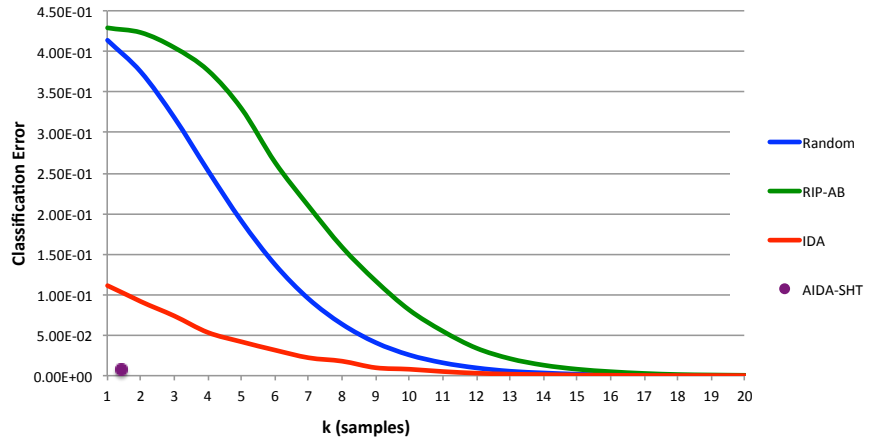


(b)

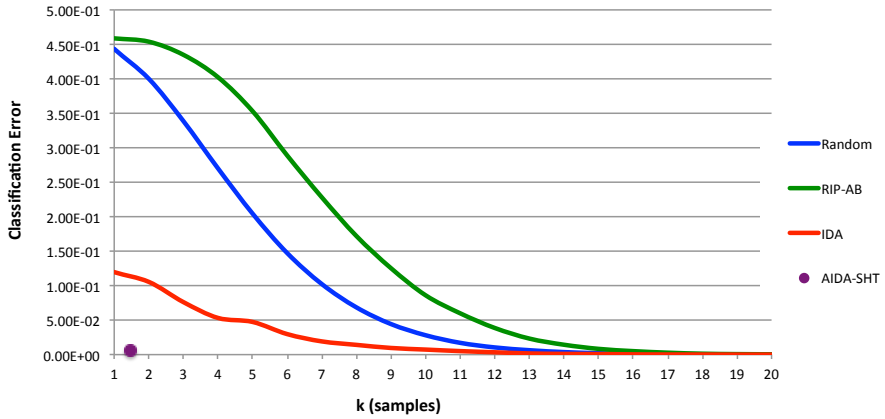


(c)

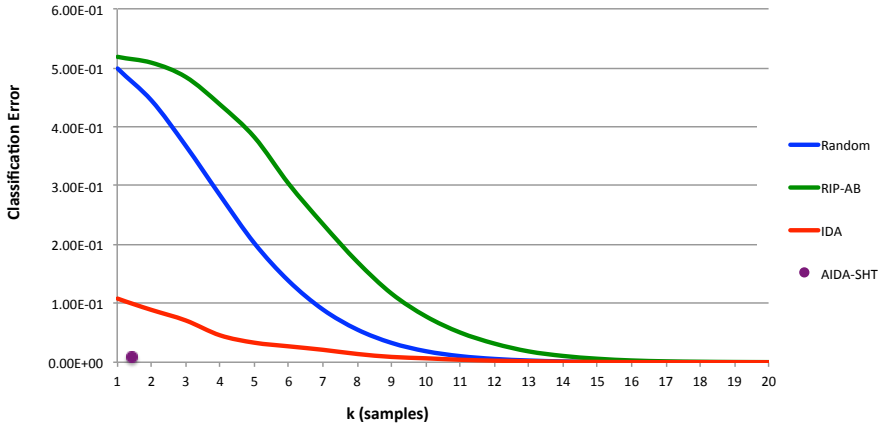
Fig. 28. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [46\ 62)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

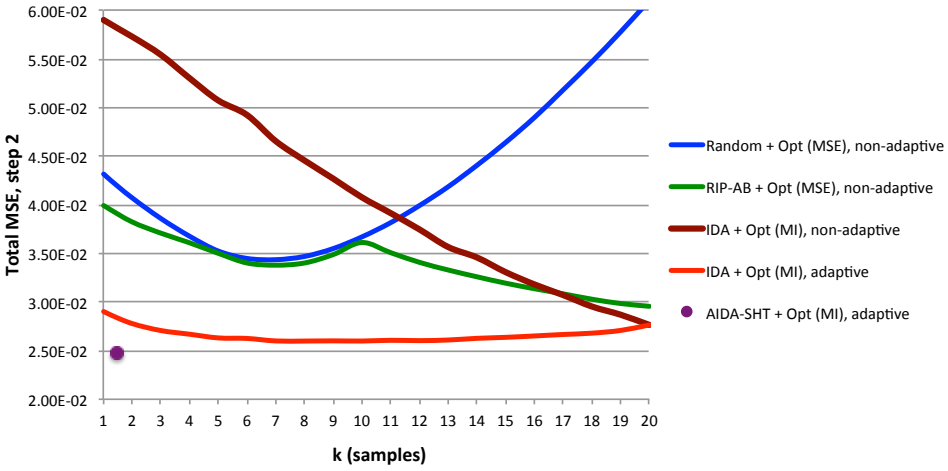


(b)

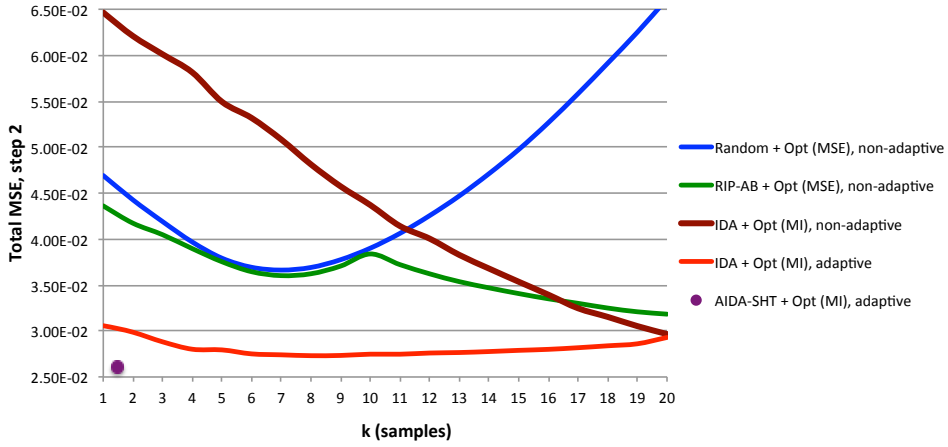


(c)

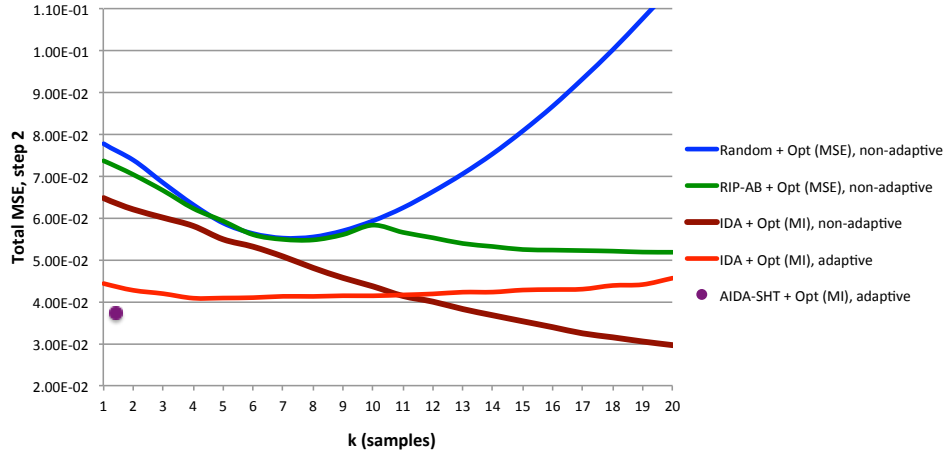
Fig. 29. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [62\ 78)$. a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)

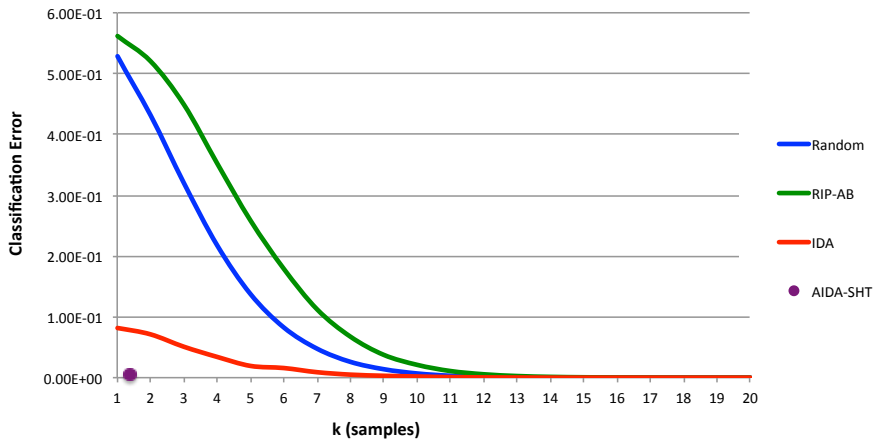


(b)

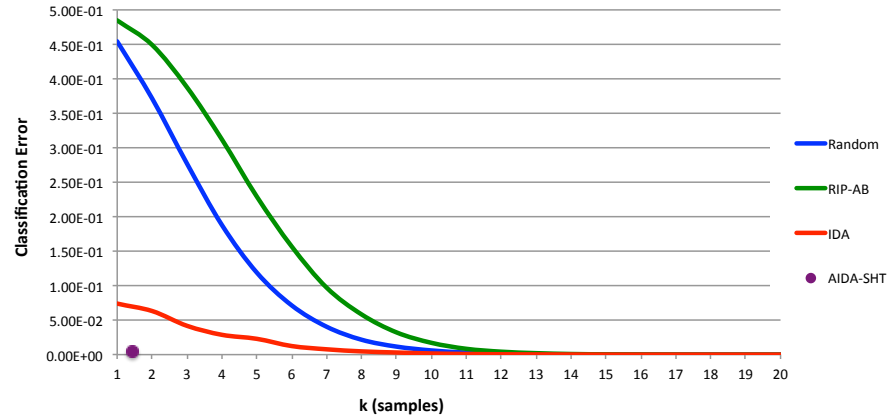


(c)

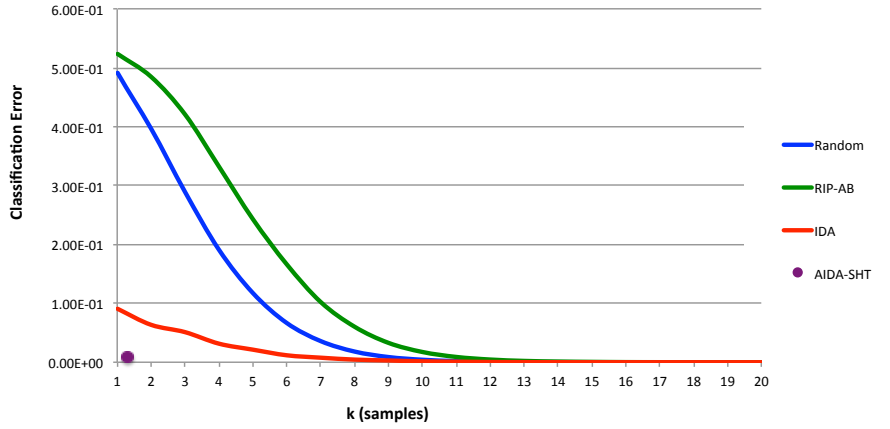
Fig. 30. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [62\ 78)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

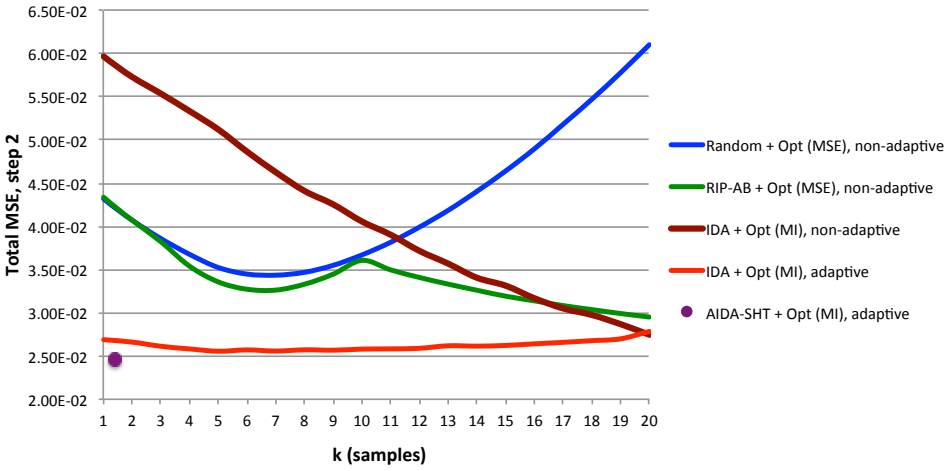


(b)

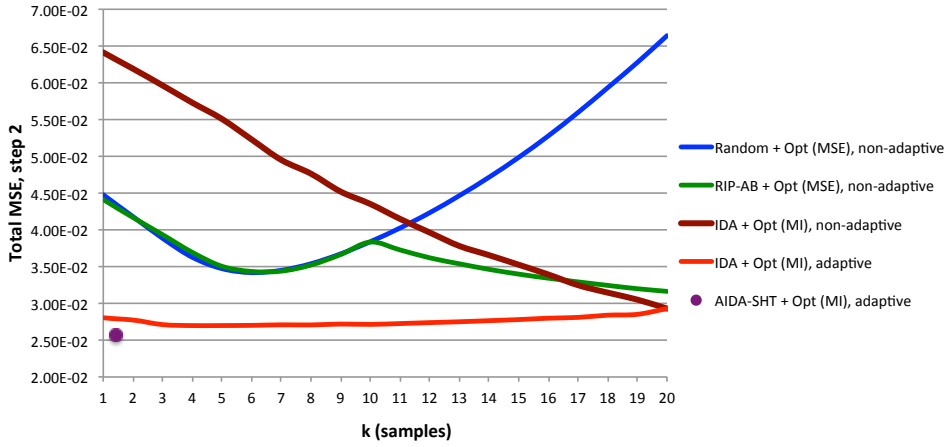


(c)

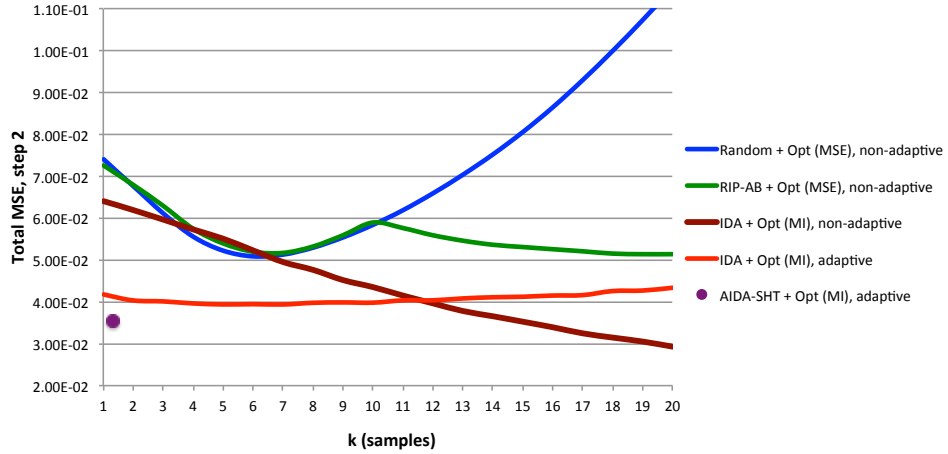
Fig. 31. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)

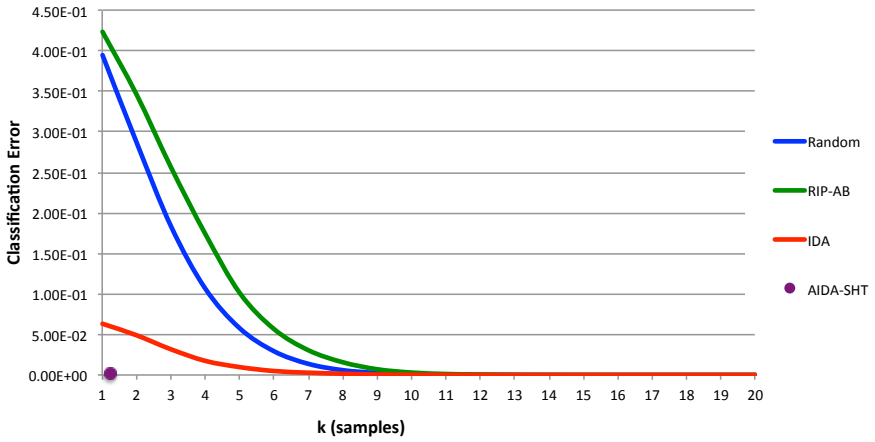


(b)

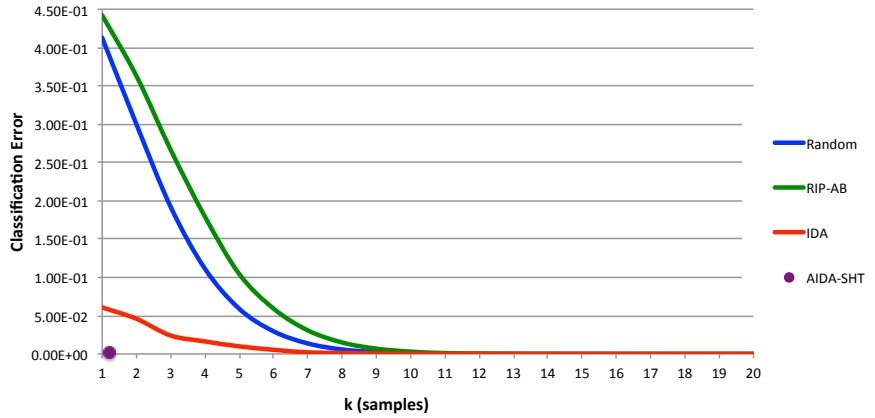


(c)

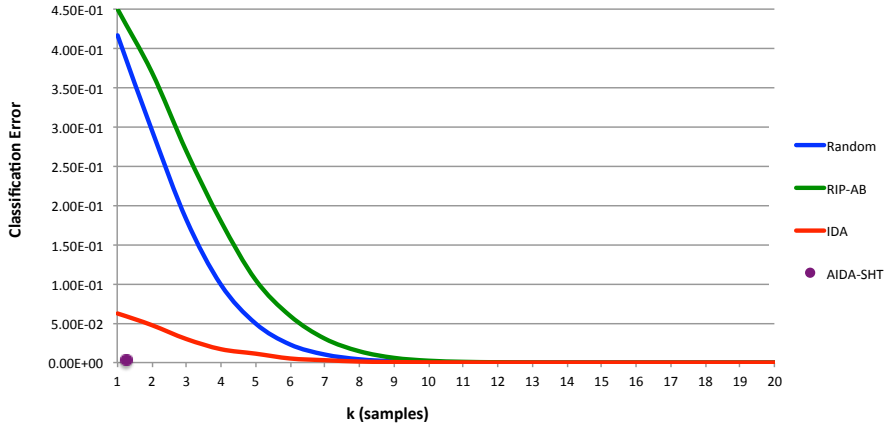
Fig. 32. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [78\ 94)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

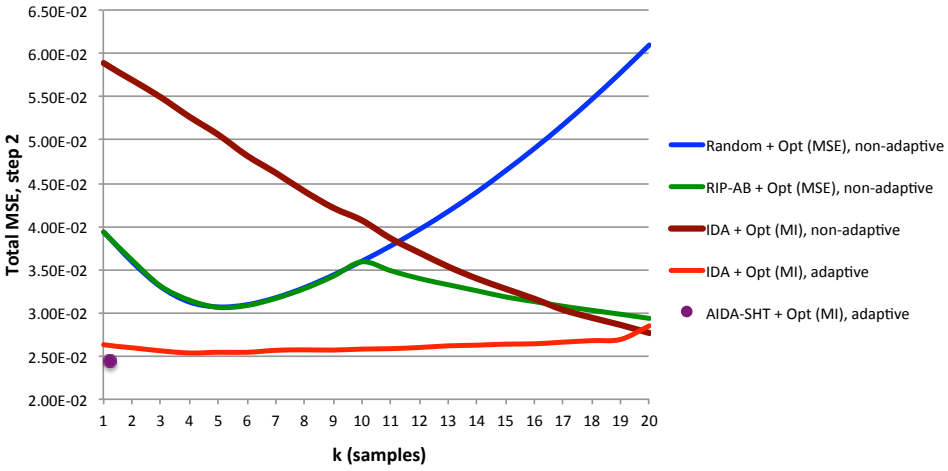


(b)

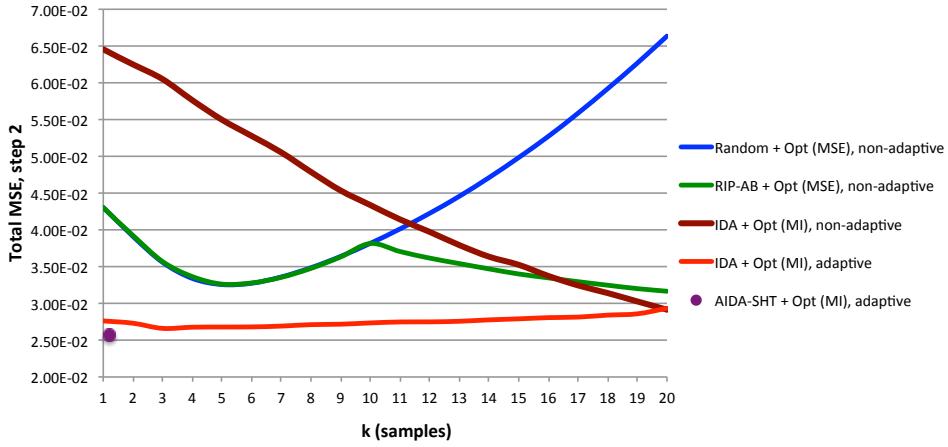


(c)

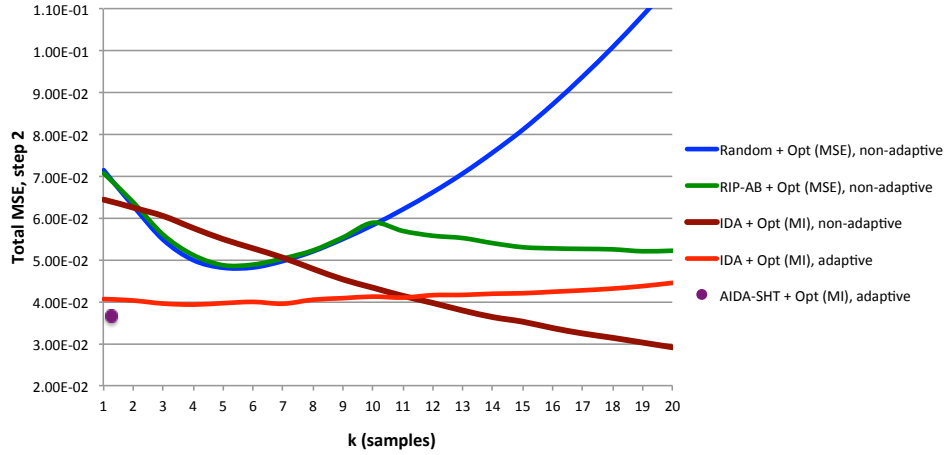
Fig. 33. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)

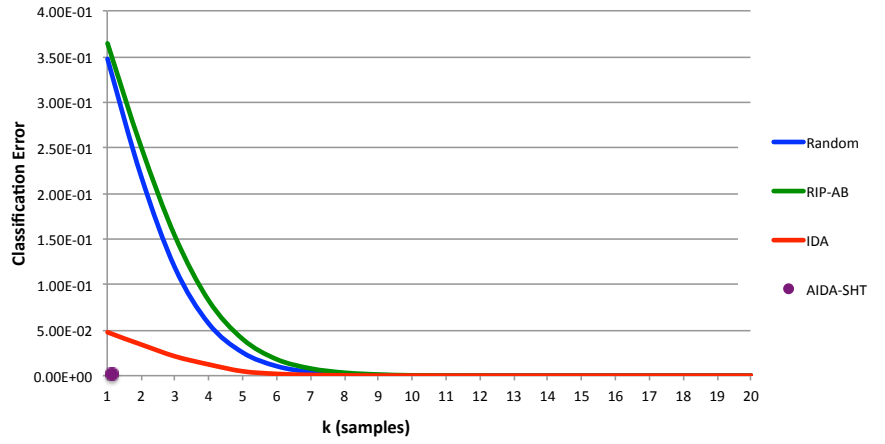


(b)

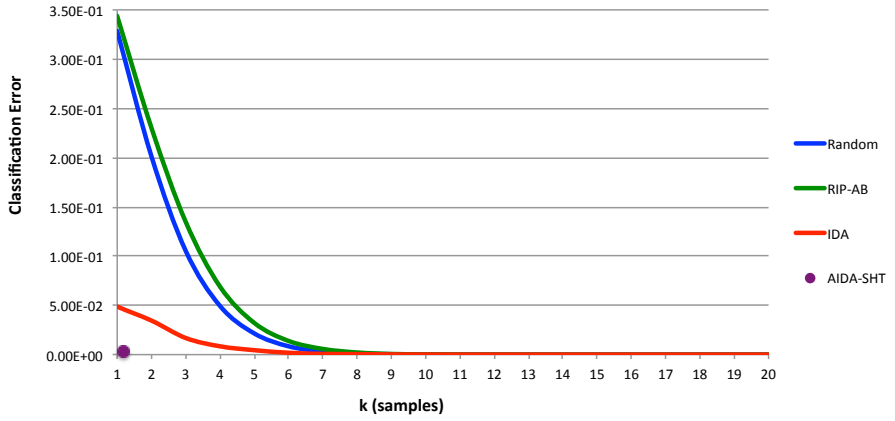


(c)

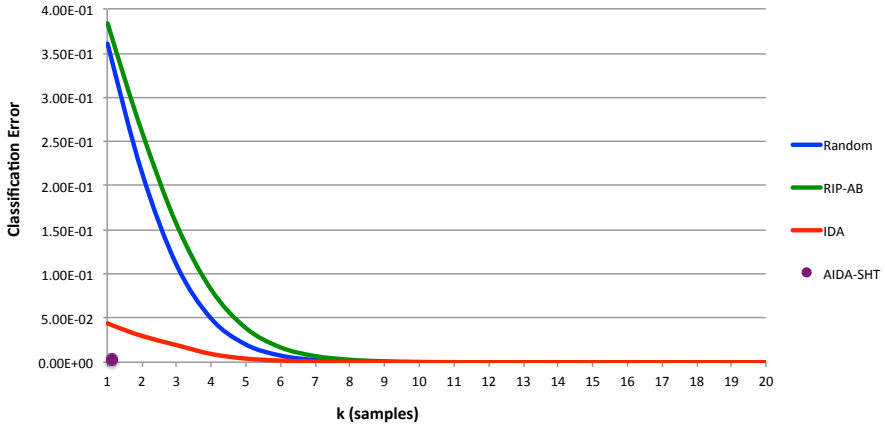
Fig. 34. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [94\ 110)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

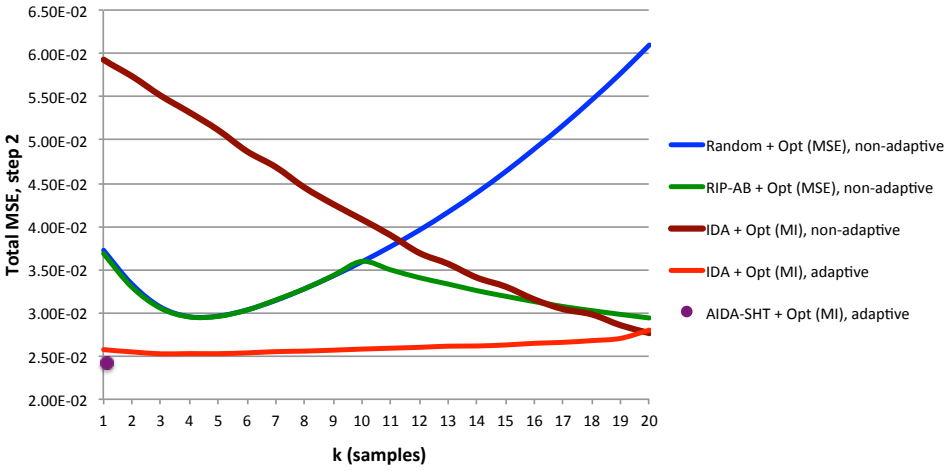


(b)

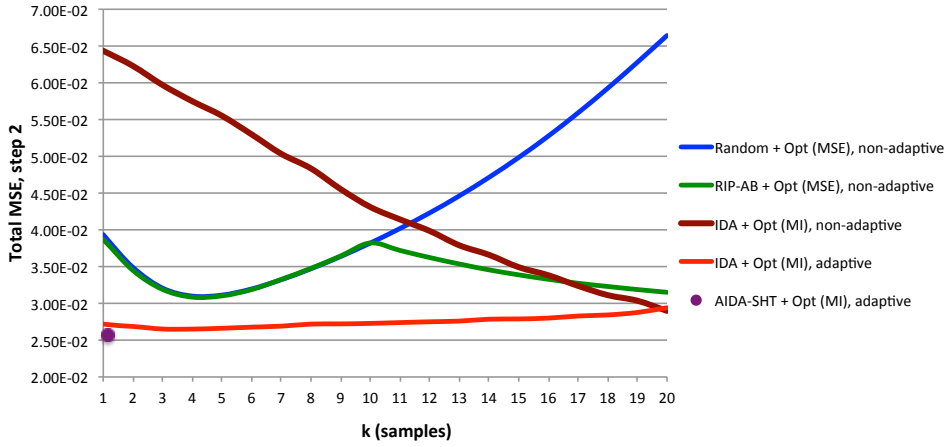


(c)

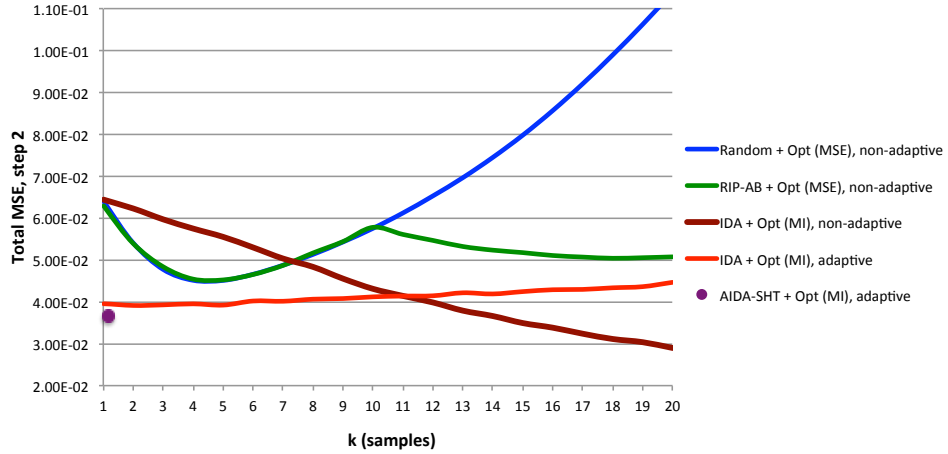
Fig. 35. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [110\ 126)$. a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)

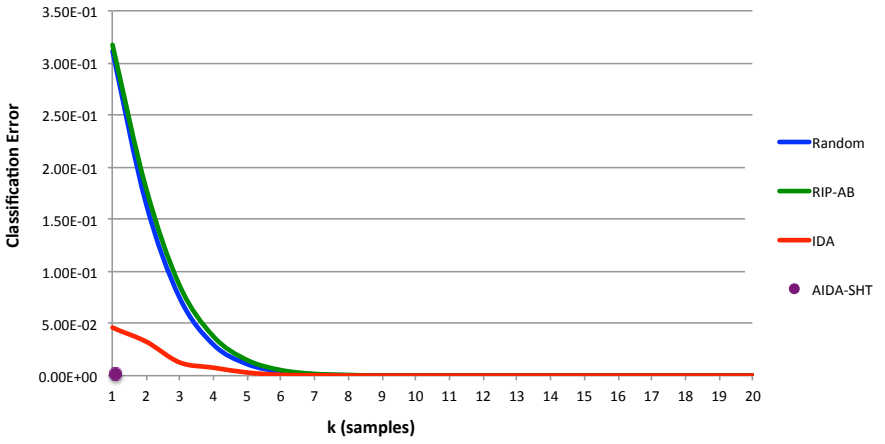


(b)

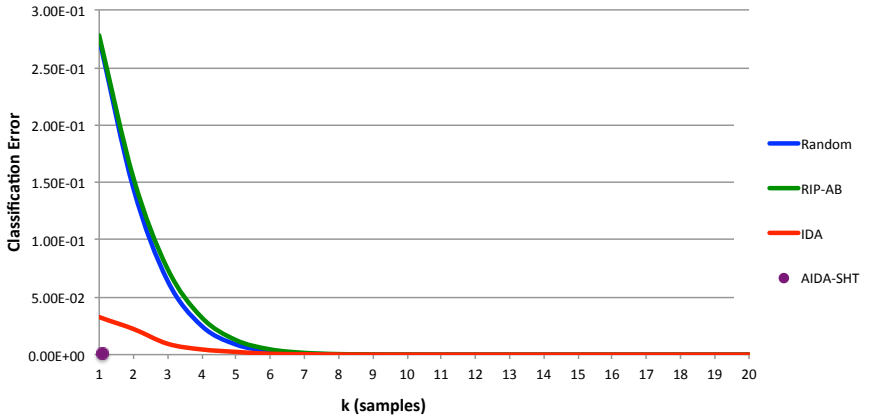


(c)

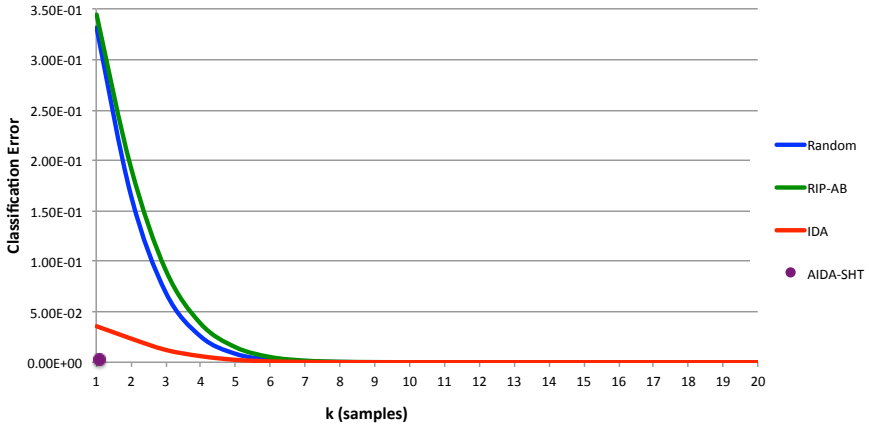
Fig. 36. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [110\ 126)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

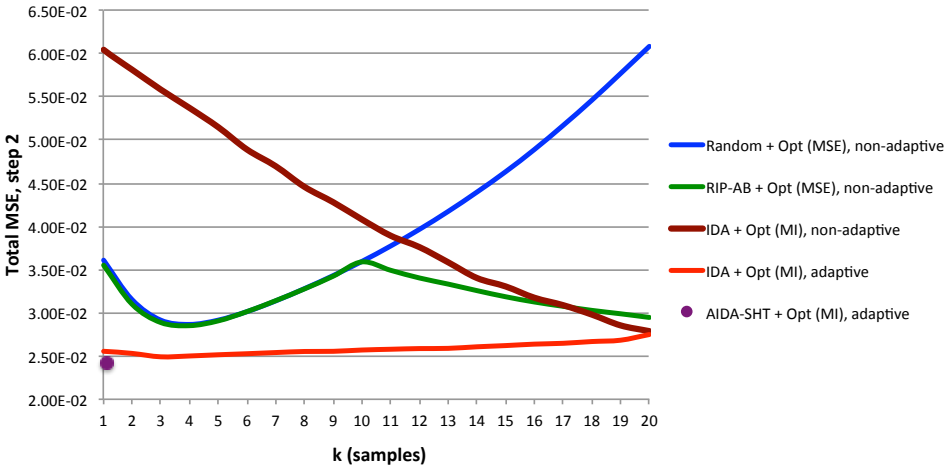


(b)

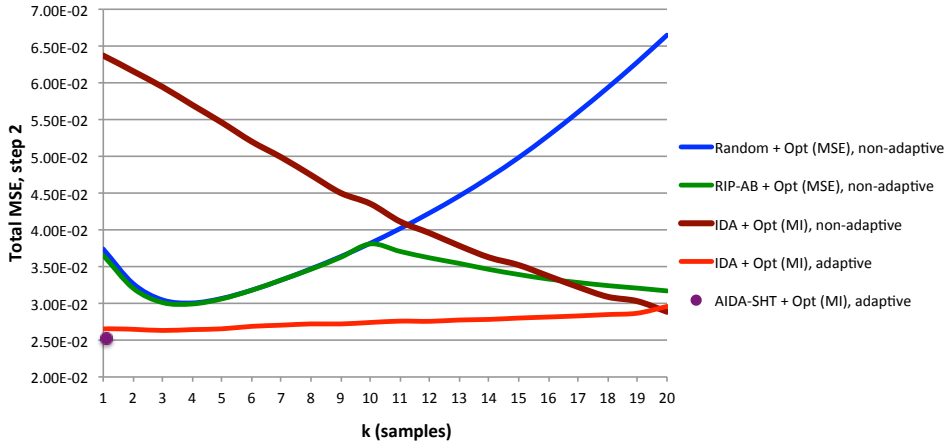


(c)

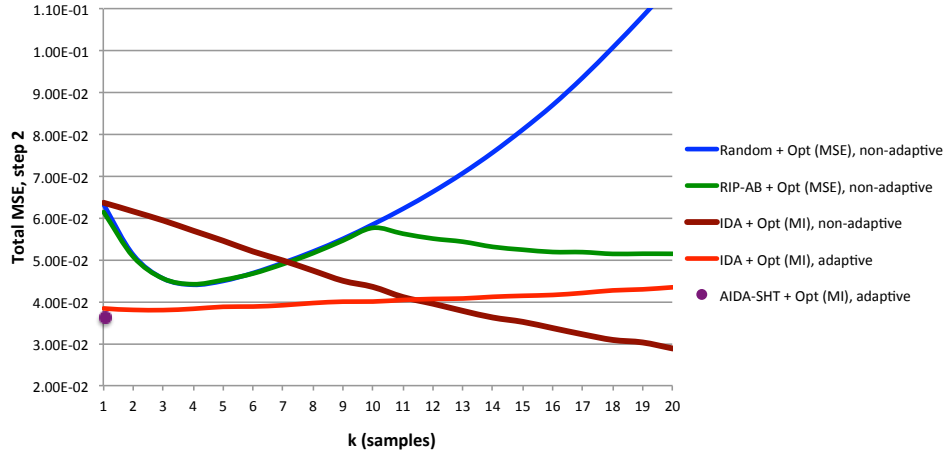
Fig. 37. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in$ [126 142). a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)

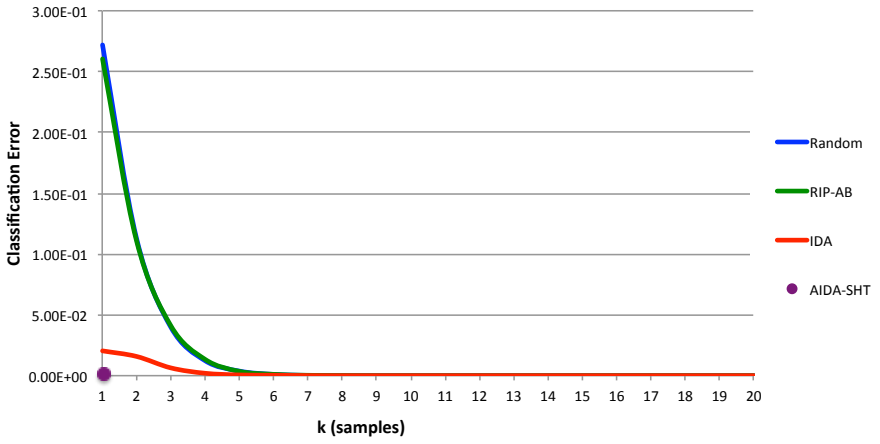


(b)

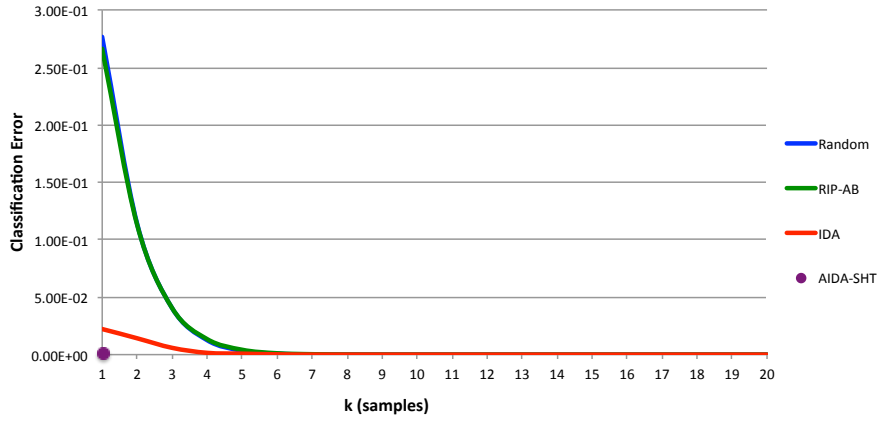


(c)

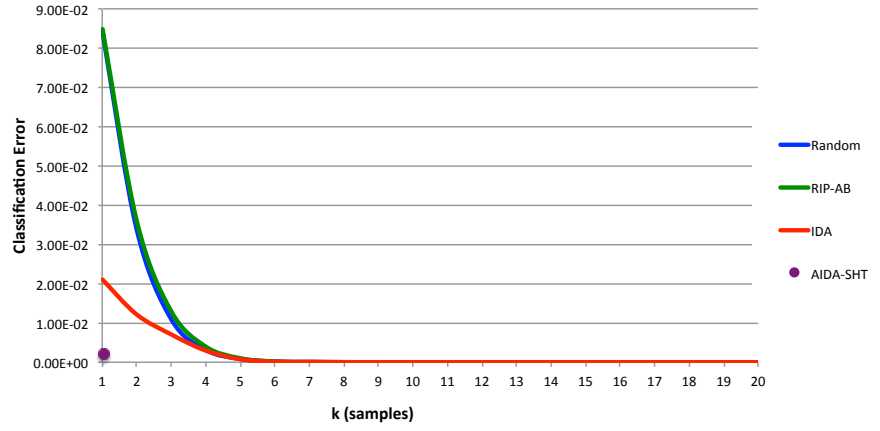
Fig. 38. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [126\ 142)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.



(a)

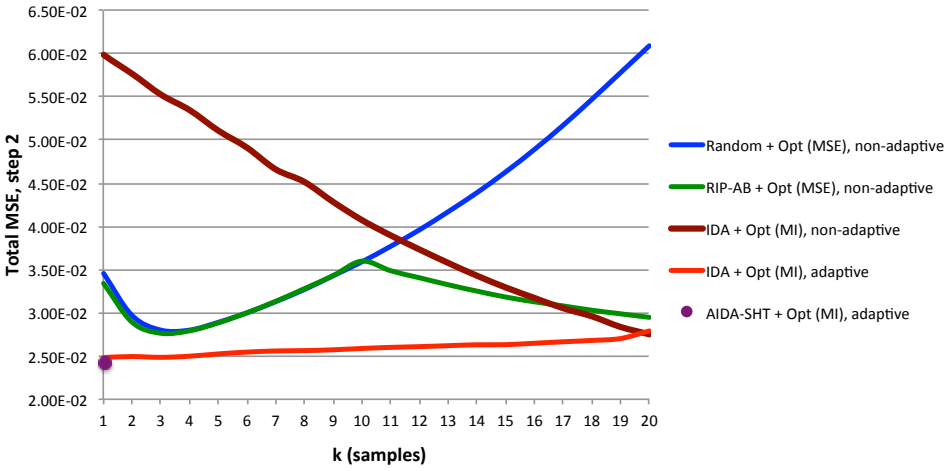


(b)

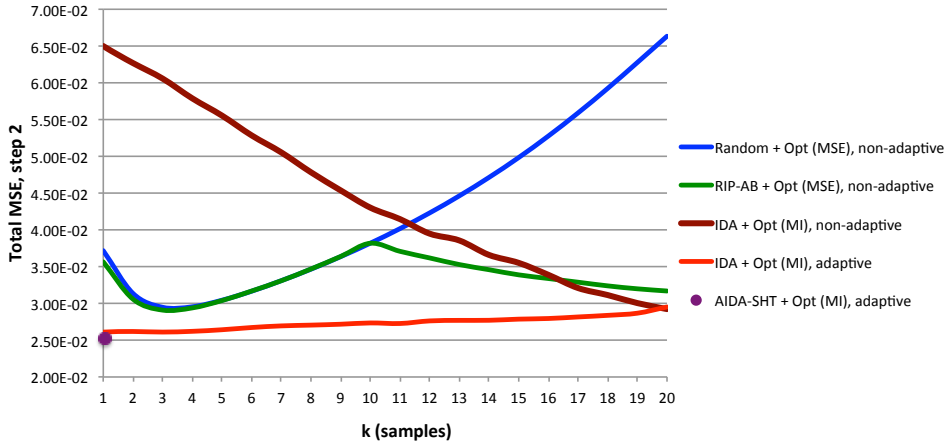


(c)

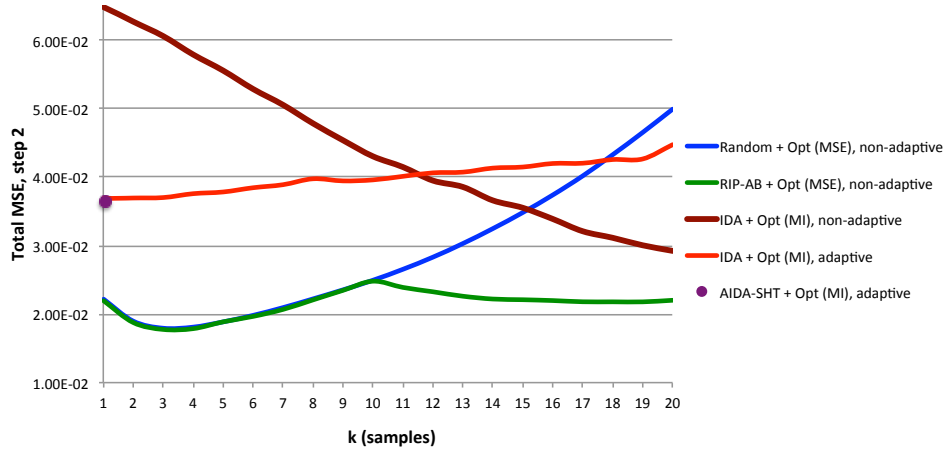
Fig. 39. Classification accuracy (step 1) synthetic signals of dimension 100 (CS to 20 samples) BD $\in [142 \ +\infty)$. a) No noise, b) SNR of 40 dbs, c) SNR of 31 dbs.



(a)



(b)



(c)

Fig. 40. MSE (step 2) reconstructed synthetic signals of dimension 100 (CS to 20 samples) BD $\in [142 \ +\infty)$. a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.

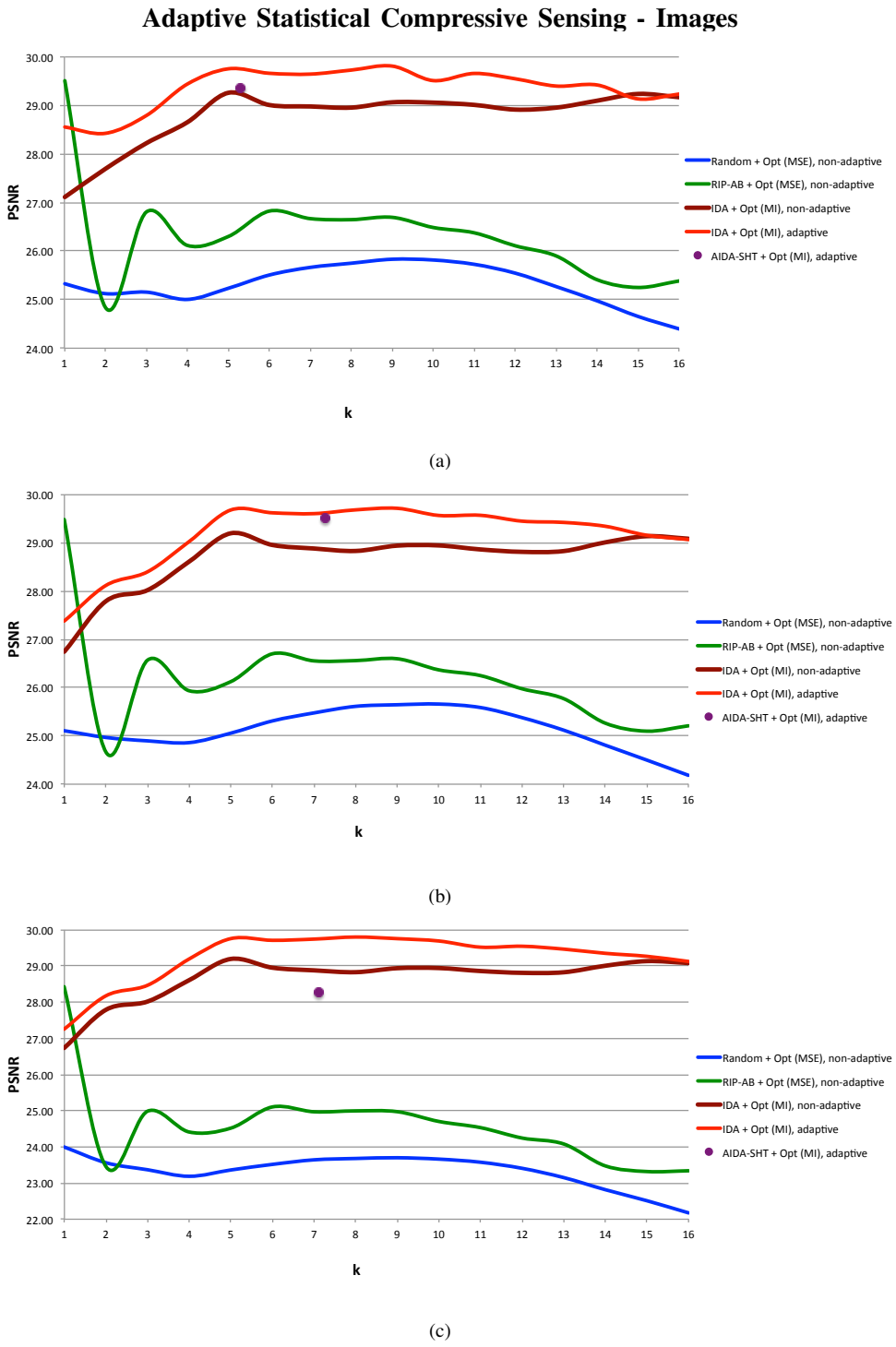


Fig. 41. PSNR (step 2) reconstructed natural images, non-overlapping patches of size 8×8 (CS to 16 samples). a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.

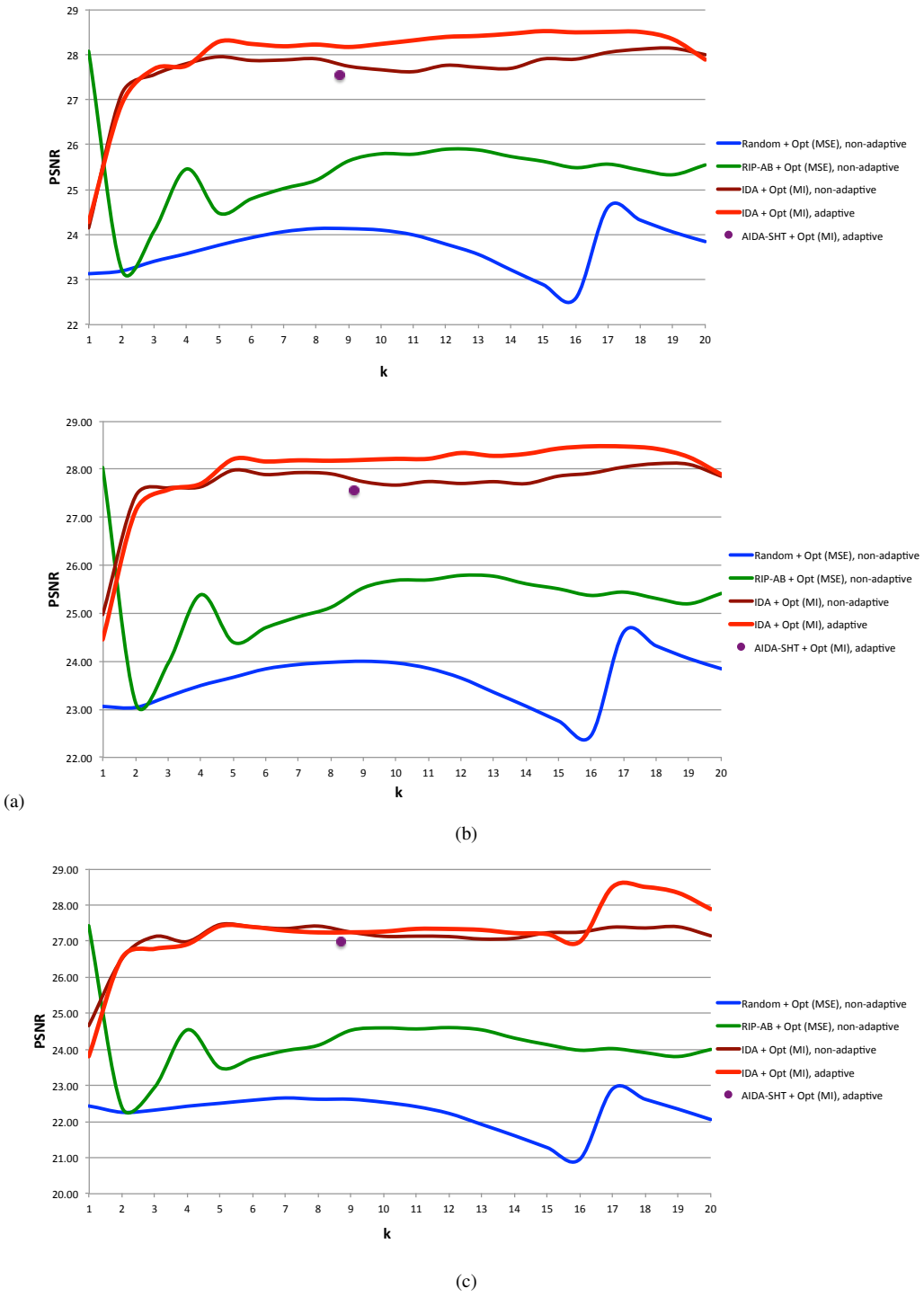


Fig. 42. PSNR (step 2) reconstructed natural images, non-overlapping patches of size 10×10 (CS to 20 samples). a) No noise, b) SNR of 40 dbs, c) SNR of 30 dbs.

(a) (b)

(c) (d)

(e) (f)

Fig. 43. Reconstructed image from non-overlapping patches of size 6×6 (CS to 6 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (20.6 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (22.3 dbs), d) IDA + Optimum (MSE) non-adaptive (23.3 dbs), e) IDA + Optimum (MI) adaptive (25.2 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (26.2 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 44. Reconstructed image from non-overlapping patches of size 6×6 (CS to 6 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (21.3 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (23.3 dbs), d) IDA + Optimum (MSE) non-adaptive (23.9 dbs), e) IDA + Optimum (MI) adaptive (25.5 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (26.5 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 45. Reconstructed image from non-overlapping patches of size 6×6 (CS to 6 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (26.5 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (28 dbs), d) IDA + Optimum (MSE) non-adaptive (29.3 dbs), e) IDA + Optimum (MI) adaptive (30.5 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (32.0 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 46. Reconstructed image from non-overlapping patches of size 8×8 (CS to 16 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (22.21 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (23.32 dbs), d) IDA + Optimum (MSE) non-adaptive (26.01 dbs), e) IDA + Optimum (MI) adaptive (27.22 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (27.17 dbs)

(a) (b)

(c) (d)

(e) (f)

Fig. 47. Reconstructed image from non-overlapping patches of size 8×8 (CS to 16 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (22.9 dbs), c) RIP-AB - Optimum (MSE) non-adaptive (24.3 dbs), d) IDA + Optimum (MSE) non-adaptive (27.0 dbs), e) IDA + Optimum (MI) adaptive (27.5 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (27.4).

(a) (b)

(c) (d)

(e) (f)

Fig. 48. Reconstructed image from non-overlapping patches of size 8×8 (CS to 16 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (28.3 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (29.6 dbs), d) IDA + Optimum (MSE) non-adaptive (31.9 dbs), e) IDA + Optimum (MI) adaptive (32.9 dbs)), and f) AIDA-SHT + Optimum (MI) adaptive (32.7).

(a) (b)

(c) (d)

(e) (f)

Fig. 49. Reconstructed image from non-overlapping patches of size 8×8 (CS to 16 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (29.6 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (31.1 dbs), d) IDA + Optimum (MSE) non-adaptive (33.1 dbs), e) IDA + Optimum (MI) adaptive (33.7 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (33.4 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 50. Reconstructed image from non-overlapping patches of size 10×10 (CS to 20 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (20. 7dbs), c) RIP-AB + Optimum (MSE) non-adaptive (21.3 dbs), d) IDA + Optimum (MSE) non-adaptive (25.6 dbs), e) IDA + Optimum (MI) adaptive (26.1 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (25.4 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 51. Reconstructed image from non-overlapping patches of size 10×10 (CS to 20 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (21.9 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (22.5 dbs), d) IDA + Optimum (MSE) non-adaptive (25.9 dbs), e) IDA + Optimum (MI) adaptive (26.2 dbs), and f) AIDA-SHT + Optimum (MI) adaptive (25.7 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 52. Reconstructed image from non-overlapping patches of size 10×10 (CS to 20 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (26.8 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (27.3 dbs), d) IDA + Optimum (MSE) non-adaptive (31.1 dbs), e) IDA + Optimum (MI) adaptive (31.3 dbs), and f) AIDA-SHT+ Optimum (MI) adaptive (30.2 dbs).

(a) (b)

(c) (d)

(e) (f)

Fig. 53. Reconstructed image from non-overlapping patches of size 10×10 (CS to 20 samples) using the following two-step protocols: a) Original, b) Random + Optimum (MSE) non-adaptive (28.8 dbs), c) RIP-AB + Optimum (MSE) non-adaptive (28.4 dbs), d) IDA + Optimum (MSE) non-adaptive (31.6 dbs), e) IDA + Optimum (MI) adaptive (31.5 dbs), and f) AIDA-SHT+ Optimum (MI) adaptive (31.2 dbs).